目　录

第一章　引　　言

1.1　绪　　论

什么是统计？统计就是收集及分析数据，并由此作出推断的科学．统计要从数据出发建立模型，这叫归纳(induction)；建立模型之后，要用它来进行推断，这叫演绎(deduction)．和以演绎为主并基于公理系统的数学不一样，统计是基于数据的，其数学基础是概率论．由于现实世界的多样性，在统计中不存在完美的模型．任何一个由数据归纳出来的模型往往要再回到实际中对其检验，并用新的数据对之进行修正．这种反复的认识及再认识的思想方法是统计的一个突出特点．数学是一个可以独立存在的逻辑体系．而对于统计来说，离开了应用，就没有存在的必要．

一般经典的数理统计教科书的主要部分是由估计和检验两大部分组成．在那里，往往假设产生数据的总体分布的形式是已知的．所不能确定的是数量有限的一些参数值，而所要做的就是对这些参数进行检验或估计．但是实践中，在没有足够证据时，去假设一个总体有某种分布形式，并进行参数估计或检验是不负责的，结果是不可靠的，甚至是灾难性的．

非参数统计就是在对总体分布形式不了解时进行推断的统计方法．这里对于总体分布不作或只作一点诸如对称性之类的简单假设．虽然不知道分布的形式，我们总可以把数据按大小排队而使每个数据都有自己的"地位"，我们称之为秩(rank)．大小为 n 的样本产生了 n 个秩．这样，问题就简化为对这些秩的研究了．幸运的是，这些秩及由其产生的一些统计量的性质和分布是可以得到的，并且与原来的总体分布无关(distribution-free)．除了与秩有关的方法之外，还有其它一些非参数方法．非参数方法有相当好的稳健性(后面要介绍)，计算简单，处理问题广泛，并且在多数分布未知的情况下比参数方法更有效．但也应指出：虽然参数方法有局限性，但在总体分布已知时，它比非参数方法利用更多的样本中的信息，因而就更有效．

本章介绍一些为学习后面章节所需要的基本的统计和概率知识．如已熟悉，可略过不看．第四节之后的部分最好在用到时再看．一些概念，如完全估计量和相容估计量等对初学者或非数理统计方向的读者也可略去不看．

1.2　估计和检验

1.2.1　点估计和区间估计

假定我们掷一枚硬币 n 次，得到 S 次正面，需要估计出现正面的概率 p．由直观，我们可用

$\frac{S}{n}$ 来估计 p. 当然，你可以用任何可以想象的其它方法来估计 p. 这样就产生了一个评价估计好坏的标准问题. 所谓"好"和"坏"，其实只是相对于你的要求而言. 在数理统计课程中已引进了各种不同的标准. 本节仅就本书中要用到的标准作一回顾.

要估计上述的概率 p，绝不能只掷一次硬币. 我们希望在大量试验中，估计量的平均值尽可能地接近所要估计的真值. 这就产生了无偏估计量(unbiased estimator) 的概念. 假设有样本 $X_1, X_2, \cdots, X_n$. 它们的总体分布(函数) 为 $F(x,\theta)$，而 θ 为要估计的参数. 如果我们选定的对 θ 的估计量是 $T(X_1,\cdots,X_n)$(注意，它是样本数据 $X_1,\cdots,X_n$ 的一个函数或统计量，与参数 θ 无关)，在满足

$$E_\theta(T(X_1,\cdots,X_n)) = \theta$$

时，我们称 $T \equiv T(X_1,\cdots,X_n)$ 为 θ 的一个无偏估计量，这里 $E_\theta(\cdot)$ 表示基于 $F(x,\theta)$ 的期望.

我们可以把掷硬币看成是 n 个独立的 Bernoulli 试验，即 S 服从二项分布：$S \sim b(n,p)$. 所以有

$$E\left(\frac{S}{n}\right) = p$$

也就是说，刚才选的对 p 的估计 $\frac{S}{n}$ 是无偏的. 注意，无偏估计可能不唯一，当然和任何其它种类的估计一样，它有它的缺点. 如果有两个统计量 T_1 和 T_2 为参数 θ 的无偏估计，我们自然要选择其方差小的，因为方差越小，统计量的可能值的分散程度越小. 一般来说，我们希望均方误差 $E(T-\theta)^2$ 越小越好. 如果在所有无偏估计中，估计量 T 使均方误差(对无偏估计，这就是方差) 最小，则称 T 为一致最小方差无偏估计(uniformly minimum variance unbiased estimator——UMVUE).

在用一个统计量 $T(X_1,\cdots,X_n)$ 估计参数 θ 时，我们当然要求这个统计量要尽量用到样本中的全部信息，在统计上，称这种统计量为充分的. 确切地说，如果在给定 $T(X_1,\cdots,X_n) = t$ 下，$(X_1,\cdots,X_n)$ 的条件分布与 t 无关，则称 $T(X_1,\cdots,X_n)$ 是分布族 $\{F(x,\theta):\theta \in \Theta\}$ 的充分统计量.

既然 UMVUE 是参数 θ 的一个好的估计，那么它是不是唯一的？为了解决这一问题，又引进了统计上另一个重要的概念 —— 完全统计量. 确切地说，对于分布族 $\{F(x,\theta):\theta \in \Theta\}$，如任给满足

$$E_\theta g(T) = 0, \quad \forall\, \theta \in \Theta$$

的函数 $g(\cdot)$，都有 $P_\theta(g(T) = 0) = 1$，则称统计量 $T(X_1,\cdots,X_n)$ 的导出分布族是完全的.

在 Bernoulli 试验中，因为

$$\lim_{n\to\infty} \mathrm{Var}\left(\frac{S}{n}\right) = \lim_{n\to\infty} E\left(\frac{S}{n} - p\right)^2 = 0$$

则对任意的 $\varepsilon > 0$ 有

$$\lim_{n\to\infty} P\left(\left|\frac{S}{n} - p\right| > \varepsilon\right) = 0$$

直观上，随着试验次数 n 的增加，估计值 $\frac{S}{n}$ 与实际的参数值 p 应更接近. 一般来说，如对任意 $\varepsilon > 0$，参数 θ 的估计量 $T(X_1,\cdots,X_n)$ 满足

$$\lim_{n\to\infty} P(|T(X_1,\cdots,X_n) - \theta| > \varepsilon) = 0$$

则称 $T(X_1,\cdots,X_n)$ 为 θ 的相容(或相合)估计量(consistent estimator). 注意,相容性是一个大样本性质,在固定的小样本情况,应谨慎对待. 有时,一个相容统计量会没有任何实际意义.

如果取 $T=T(X_1,\cdots,X_n)$ 作为 θ 的一个估计,我们能用它来估计 θ 的一个可能的范围或其可能的上下界. 一个常用的范围的形式为 $T(X_1,\cdots,X_n)\pm a$. 当然,因为 T 是个随机变量,所以我们只能说由它导出的区间(置信区间)以某概率(置信度)覆盖参数 θ. 一般地说,如果 $[T_l,T_u]$ 是由一对统计量 $T_l,T_u(T_l\leqslant T_u)$ 所组成的随机区间,如对所有的 θ 有

$$P_\theta(T_l\leqslant\theta\leqslant T_u)=1-\alpha$$

则称 $[T_l,T_u]$ 为 θ 的置信度为 $1-\alpha$ 的置信区间(confidence interval). 这里 $P_\theta(\cdot)$ 表示当 θ 为真实参数值时的概率. 换言之,我们以 $100(1-\alpha)\%$ 的概率或置信度(confidence level)保证 $[T_l,T_u]$ 覆盖 θ.

1.2.2 假设检验

如果在上面掷硬币的试验中,我们怀疑硬币的均匀性,即怀疑是否 $p=\frac{1}{2}$. 我们就要对原假设(null hypothesis) $H_0: p=\frac{1}{2}$ 进行检验. 备择假设(alternative hypothesis)可为 $p\neq\frac{1}{2}$, $p<\frac{1}{2}$ 及 $p>\frac{1}{2}$ 三者之一. 如果备择假设用 $p\neq\frac{1}{2}$,则称检验是双边的. 如备择假设用另外两个之一,则称检验是单边的.

对原假设进行检验的结果只能是下列两个决策之一:1. 拒绝原假设 H_0;2. 不能拒绝原假设 H_0. 有些作者用"接受备择假设"来代替第 2 个决策,这是不对的. 因为在检验中,我们一直在原假设条件下进行概率运算,在原假设不对时,没有任何理由来"接受"备择假设. 我们尊重他人基于历史原因的选词,但为了科学的准确性及避免逻辑混乱,我们不主张用"接受备择假设"的说法.

上面的原假设只包含一个点,称为简单假设(simple hypothesis). 一般地,假定 Θ 为所有可能的参数值 θ 的集合. 原假设为 $\theta\in\Theta_0$,备择假设为 $\theta\in\Theta_1$. 而 $\Theta_0\subset\Theta$, $\Theta_1\subset\Theta$ 及 $\Theta_0\cap\Theta_1=\varnothing$. 当 Θ_0 包含多于一个点时,称检验为复合假设(composite hypothesis). 注意,在简单假设下,分布被唯一确定. 而在复合假设情况则不尽然.

在检验中,我们需要选择一个检验统计量(test statistic): $T\equiv T(X_1,\cdots,X_n)$. 因为检验统计量完全确定了检验的性质,所以,检验统计量也称为检验. 在原假设成立时,它的可能值只以很小的概率属于某个范围,比如集合 W. 如果事件 $(T\in W)$ 的确发生了,它在原假设下是一个小概率事件. 换句话说,原假设有问题,应该拒绝. 这时, W 称为拒绝域(rejection region 或 critical region). 如果事件 $(T\notin W)$ 发生了,则我们没有理由拒绝原假设. 当 W 是诸如 $(-\infty,c]$ 或 $[c,\infty)$ 一类的区间时, $T\in W$ 等价于 $T\leqslant c$ 或 $T\geqslant c$. 这时称 c 为临界值(critical value). 在决策中,我们可能会犯两种错误. 一种是原假设对,我们拒绝了它,这是所谓的第 I 类错误;另一种是原假设不对,但没有拒绝,即所谓的第 II 类错误. 犯这两类错误的概率分别为 $P_\theta(T\in W|\theta\in\Theta_0)$ 和 $P_\theta(T\notin W|\theta\in\Theta_1)$. 人们自然会希望这两个概率越小越好,但在样本给定之后不可能两全其美. 通常是先限制第 I 类错误概率不大于预先给定的概率 $0<\alpha<1$,它被称为显著性水平(level of significance)或检验水平(size of test). 即对任意的 $\theta\in\Theta_0$,

$$P_\theta(T\in W)\leqslant\alpha.$$

在此条件下，选择合适的检验统计量使犯第 II 类错误的概率尽可能地小，即使 $P_\theta(T \in W|\theta \in \Theta_1)$ 尽可能地大. 我们称 θ 的函数 $\beta(\theta) = P_\theta(T \in W)$ 为势(函数)(power function). 显然当 $\theta \in \Theta_0$ 时，$\beta(\theta)$ 是犯第 I 类错误的概率. 而当 $\theta \in \Theta_1$ 时，$1-\beta(\theta)$ 是犯第 II 类错误的概率. 上面的限制第 I 类错误概率条件可写成

$$\sup_{\theta \in \Theta_0} \beta(\theta) \leqslant \alpha$$

注意，势函数实际上也依赖于检验 T 的选择，我们可记它为 $\beta(\theta, T)$. 如果一个水平 α 的检验 T^* 使得对于所有的水平 α 的检验 T 及所有的 $\theta \in \Theta_1$ 有

$$\beta(\theta, T^*) \geqslant \beta(\theta, T)$$

则称检验 T^* 是一致最优势的(uniformly most powerful——UMP). 因为人们总希望在水平 α 尽量小的时候拒绝原假设. 举例说，如果我们可以在 $\alpha = 0.01$ 拒绝，当然也可以在 $\alpha = 0.05$ 拒绝；但总是选小的 α 以证明我们拒绝得有道理. 因此，在实践及各种计算机软件中，人们并不预先指定水平的值，而是很方便地利用由数据产生的下面定义的 p 值. 在取得了 $X_1, \cdots, X_n$ 的观察值 $x_1, \cdots, x_n$ 之后，我们称概率

$$P_\theta(T(x_1, \cdots, x_n) \in W | \theta \in \Theta_0)$$

为该检验的 p 值(p-value) 或观察水平(observed size) 或显著概率(significance probability). 对于任何大于 p 值的水平，人们可以拒绝原假设，但不能在任何小于它的水平下拒绝原假设. p 值是使人们可以拒绝原假设的最小水平.

例 1.1 假设在 $n = 10$ 次掷硬币的试验中，共出现正面 $S = 3$ 次. 要检验该硬币是否均匀. 令 θ 为出现正面的概率. 原假设为 $H_0: \theta = 0.5$，而备择假设为 $H_1: \theta < 0.5$. p 值为

$$P_{0.5}(S \leqslant 3) = 2^{-10} \sum_{k=0}^{3} C_{10}^k = 0.1719$$

因此，对于所有小于 0.1719 的水平，我们不能拒绝原假设.

1.2.3 稳健性及稳健统计

我们知道，统计就是要使所建立的模型和其所反映的现实世界尽可能地一致. 但是，不存在完美的模型，也不存在不含误差的数据. 只能希望我们的方法或模型对于有危险的误差不至于太敏感. 这就是稳健性的概念(robustness). 稳健概念实际上是针对统计中的假设过分理想化而产生的. 稳健性是非参数统计的基本特点. 但是稳健统计是介于非参数统计和经典的(参数)统计之间的一些理论的集合，它是近似半参数模型的统计. 稳键统计的目的主要有以下几条：

1. 描述出适合于大多数数据的结构；2. 找出离群值(outliers)，如果需要的话，改变我们已有的结构；3. 在不平衡的数据结构中(如在回归分析中)，发现高度有影响的数据点(leverage points)，并给出警告；4. 对假定的诸如独立性等的相关结构进行审查并改进.

实际上，对于一个不太熟悉的数据结构，很难说清哪些影响点是真正满足我们要找的模型还是纯属误差的产物. 这就要对问题的背景有所了解. 纯数学式的思维方式是行不通的.

下面给出一个例子对稳健性进行说明.

例 1.2 设 $F(x)$ 为一关于 μ 对称的连续分布函数. $X_1,\cdots,X_n$ 是服从该分布的一个样本. 我们来比较两个 μ 的估计量. 一个是样本均值 $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}x_i$, 另一个是样本中位数 X_{med}, 定义为顺序统计量的中间值, 即当 n 为偶数时它取 $\frac{\left(X_{\left(\frac{n}{2}\right)}+X_{\left(\frac{n}{2}+1\right)}\right)}{2}$, 而当 n 为奇数时它为 $X_{\left(\frac{n+1}{2}\right)}$. 这里顺序统计量 $X_{(1)}\leqslant\cdots\leqslant X_{(n)}$ 是按自小到大次序重新排列的 $X_1,\cdots,X_n$. 显然如果 $X_{(n)}$ 趋于无穷大, 则 $\overline{X}$ 也趋于无穷. 这说明 $\overline{X}$ 对个别数据的不寻常值很敏感. 而 X_{med} 则不因 $X_{(n)}$ 的异常变化而改变, 即 X_{med} 是 μ 的一个稳健估计. 我们还可看出, 虽然样本中位数具有稳健性, 但样本均值包含了更多的样本所具有的信息. 因此, 在不存在异常点时, 样本均值是更常用的.

1.3 数据初步分析

在拿到一个新的数据之后, 首先要有对该数据的直观了解. 本节介绍一些简单的数据分析, 使我们对数据的特点、大概的分布形状等有个粗略的了解, 为以后的进一步统计推断作好准备.

假定我们有三个班的 97 个学生的考试成绩表(表 1.1).

表 1.1 考试成绩

一班			二班			三班		
82	45	89	99	87	72	58	46	72
82	67	72	81	88	82	84	74	48
64	89	93	66	71	88	116	91	69
78	87	75	58	84	68	53	65	109
115	57	86	86	70	88	91	69	69
73	86	85	91	77	108	86	45	48
82	90	104	109	73	81	61	70	84
64	83	77	96	60	92	96	63	90
83	78	81	85	104	98			
96	62	77	104	57	25			
53	113	67	96	74	74			
103	39		72	96	88			
			84	62				

表中成绩是按学生姓氏笔画排列的, 人们从中并不容易一眼看出该数据的特征. 下面将对它进行初步的分析.

1.3.1 直方图

最常用的一个表现数据的方法是直方图(histogram). 它通常把数据的值域分成若干相等区间,于是数据就按区间分成若干组,每组作成一个矩形,其高和该组中数据的多少成比例,其底为所属区间. 这些矩形就是直方图,它给数据的分布一个直观的形象. 图 1.1 就是表 1.1 的数据的直方图. 这里数据被分成 10 个区间,并形成 10 个矩形. 比如分数 40—49 有 5 个人,相应地形成高为 5(至多乘一常数),宽为 10 的位于该区间的矩形.

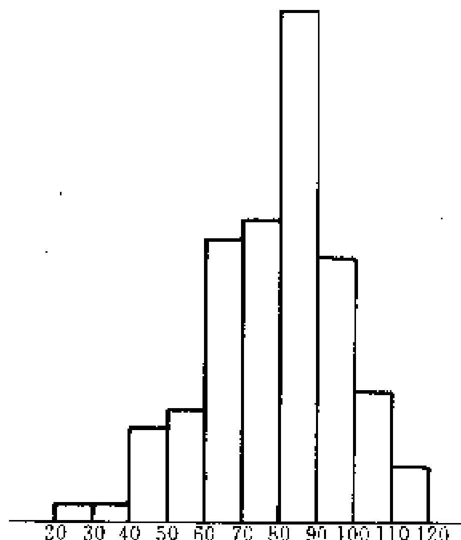


图 1.1 表 1.1 数据的直方图

一般说来,对于观测数据 $X_1,\cdots,X_n$,选择两个适当的常数 X_0 和 $h(>0)$,把 $(-\infty,+\infty)$ 分成一些小区间 $\Delta_i=[X_0+(i-1)h,X_0+ih)$, $i=0,\pm1,\pm2,\cdots$,并以 n_i 记 $X_1,\cdots,X_n$ 落在 Δ_i 的个数. 我们以 Δ_i 为底,$\frac{n_i}{nh}$ 为高做一矩形. 对 $i=0,\pm1,\pm2,\cdots$ 而得的许多矩形就是一个直方图. 直方图的形状依赖于区间的选择. 数据的特点及画图者的观点都对此有影响.

1.3.2 茎叶图

一个茎叶图(stem-and-leaf display)和直方图类似,只不过用数据代替矩形. 具体地说,把数据按除了最后一位数之外的前面数字的异同来分组;相同的分为一组(或若干组,依具体数据情况而定). 每一组数占一行,以前面的数字作为该行的标记,放在行头;并把这些数按由小到大的顺序从上往下排,这就形成了一个"茎". 每一行则是该组的所有数据的最后一位数字的排列(通常按由小到大的顺序从左至右排列),这就是"叶子". 一组中,数据越多"叶子"越长. 这既直观,又显示了具体数据.

我们把表 1.1 中的得分作出若干茎叶图:图 1.2 是三班成绩的茎叶图(没有按大小排"叶子"). 图 1.3 是所有学生的成绩的茎叶图(每行按大小排列). 图 1.4 是二班和三班成绩的背靠背茎叶图(back to back stem and leaf display),它使这两个班的成绩共用一个茎,但两个班的"叶子"分别向上下两边排列. 从该图可看出两个班成绩的不同分布特点. 这些图的茎中的值是

不同的，但如前面提到的，若干行可共有相同的茎值. 图 1.5 和图 1.6 就是对一个班 30 个学生成绩的因不同分组而作成的两个茎叶图，其中一个茎值相应于若干行.

茎	叶					
4	6	8	5	8		
5	8	3				
6	9	5	9	9	1	3
7	2	4	0			
8	4	6	4			
9	1	1	6	0		
10	9					
11	6					

图 1.2　三班成绩的茎叶图

						9			
						9			
						8			
						8			
						8			
						8			
						7			
						7			
						6			
					8	6			
					8	6			
					7	6			
				9	7	5			
				9	5	5	9		
				9	4	4	8		
				8	4	4	6		
				7	4	4	6		
				7	3	4	6		
				6	3	3	6		
				5	2	3	6		
				4	2	2	3	9	
			8	4	2	2	2	9	
		8	8	3	2	2	1	8	
		8	7	2	1	2	1	4	
		6	7	2	1	1	1	4	6
		5	3	1	0	1	0	4	5
5	9	5	3	0	0	1	0	3	3
2	3	4	5	6	7	8	9	10	11

图 1.3　表 1.1 的茎叶图

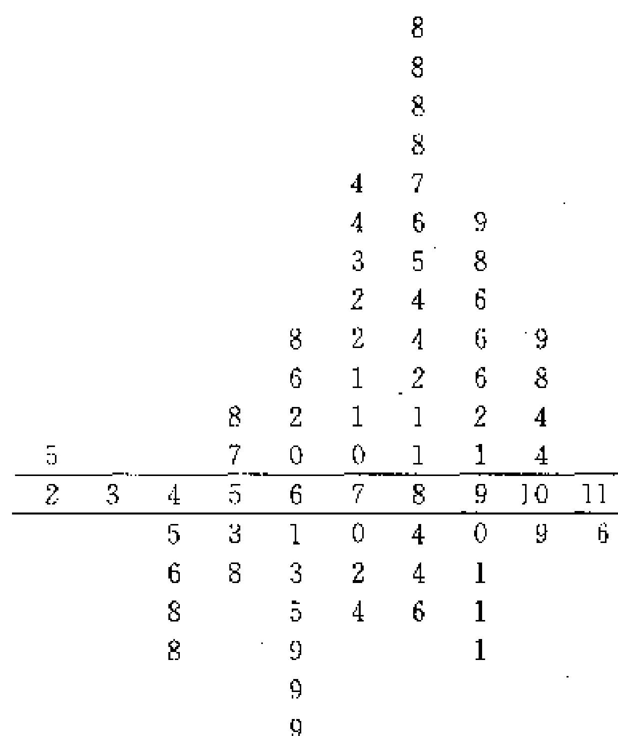


图 1.4 背靠背茎叶图

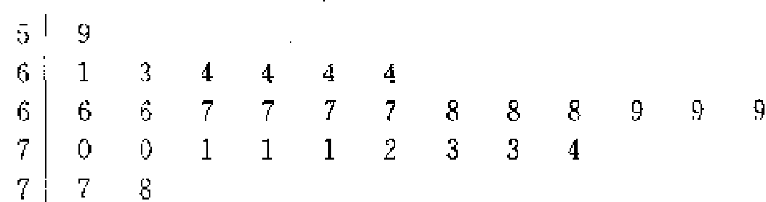


图 1.5 分两组的茎叶图

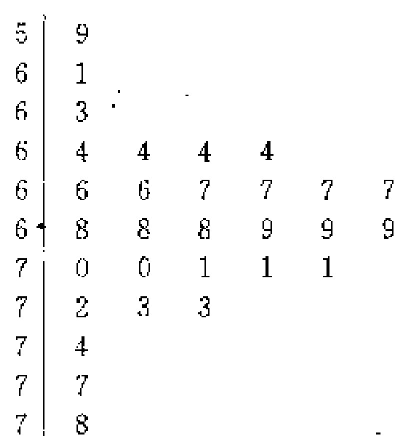


图 1.6 分五组的茎叶图

1.3.3 五数概括

直方图和茎叶图包含了大量的样本信息，但没有作任何加工或简化. 我们有时需要用少数

几个统计量来对大量的原始数据进行概括. 下面引进所谓的五数概括(five-number summaries).

有了一组数据之后,我们首先感兴趣的可能是数据的"中心". 通常人们首先想到的"中心"的度量是样本均值. 样本均值的确用得很多,但正如前面所说,样本中位数也是一个可取的关于数据"中心"的度量,它具有某种稳健性. 这一节我们就用它来度量数据的"中心".

我们引入层(depth)的概念. 如果把数据按大小次序排列(假定有 n 个数据),则最外面的两个,即最大和最小的两个数称为第一层,然后依次往里称为第二层,第三层等等. 当 n 为奇数时,最后一层(第$\frac{n+1}{2}$层)只剩一个数,即中位数;而当 n 为偶数时,最后一层(第$\frac{n}{2}$层)剩两个数,它们的平均是中位数. 我们用 μ 表示中位数,其层数用 $d(\mu)$ 表示.

例 1.3

有 8 个数:	46	48	58	72	74	84	91	116
层:	1	2	3	4	4	3	2	1

在茎叶图中,定义茎中每一数的层为该茎所对应的叶中数据的层的最大值. 中位数所在的茎的层为该叶所具有的数据数目,并用(·)表示.

例 1.4 图 1.7 为表 1.1 数据的茎叶图.

层	茎	叶												
1	2	5												
2	3	9												
7	4	5	5	6	8	8								
13	5	3	3	7	7	8	8							
28	6	0	1	2	2	3	4	4	5	6	…			
46	7	0	0	1	1	2	2	2	2	3	3	4	…	
(27)	8	1	1	1	2	2	2	2	3	3	4	4	4	4 …
24	9	0	0	1	1	1	2	3	6	6	6	…		
10	10	3	4	4	4	8	9	9						
3	11	3	5	6										

图 1.7 层及中位数

从该图亦可找到中位数. 因中位数的层数为 $d(\mu)=49$,而图中茎值为7的层为46,所以,茎值为8的叶中第三小的数为中位数($\mu=81$).

除了中位数之外,我们还对数据的分散程度感兴趣. 这里我们不考虑样本方差,而考虑极大值、极小值、上四分位数(upper quantile)Q_U 及下四分位数(lower quantile)Q_L. 四分位数的层定义为 $d(Q)=\frac{n+2}{4}$ 或 $d(Q)=\frac{n+3}{4}$ 依 n 为偶数或奇数而定. 得到了层数也就有了上下两个四分位数. 从例 1.3 的数据,易得 $Q_L=53$ 和 $Q_U=87.5$. 中位数、极值和四分位数就是我们所谓的五数概括. 数据落在 Q_L 和 Q_U 之间的概率为 0.5. 在它们之外太远的数则有可能为异常值. 记 $H=Q_U-Q_L$. 人们认为在区间($Q_L-1.5H, Q_U+1.5H$)之外的数据可看作是异常值. 如表 1.1 中的 25 可为一例.

1.3.4 盒子图

五数概括并不直观，现把这五个数画在一个图上：在 Q_L 与 Q_U 之间画一矩形盒子，在极大值与 Q_U 之间，极小值与 Q_L 之间画两线段，并在中位数处画一竖线就成了我们的盒子图(box-plot)。图 1.8 为表 1.1 数据的盒子图。

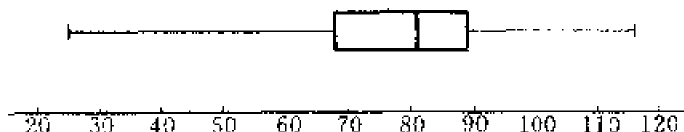


图 1.8 表 1.1 中数据的盒子图

图中矩形描述了中间的 50% 的数据；左右的水平线段代表了上下 25% 的数据的分布情况。图 1.8 显示出高低两部分数据(各占 25% 的数据)并不对称。数据在中位数 81 附近还是集中的(盒子短)。

以上这节所作的数据分析虽然很初等，但是简单明了，直观性强。它不要求数据符合任何统计模型，是获得数据之后的一种处理方法。

1.4 顺序统计量的基本性质

非参数统计的一大特点就是利用样本数据的大小关系来进行研究。因此，对在第一节已涉及的顺序统计量(order statistics)的研究构成了非参数统计的基础。设

$$X_{(1)} \leqslant X_{(2)} \leqslant \cdots \leqslant X_{(n)}$$

为样本 $X_1, \cdots, X_n$ 的顺序统计量。$X_{(i)}$ 为第 i 个顺序统计量；$X_{(1)}$ 和 $X_{(n)}$ 分别称为极小值和极大值；$X_{(n)} - X_{(1)}$ 称为极差。在第一节中，我们已用顺序统计量来表示了样本中位数。实际上，前面所讲的"五数"概括中的五数都是 p 分位数(p-quantile) 的特例。p 分位数 定义为

$$m_p = X_{([np])} + (n+1)\left(p - \frac{[np]}{n+1}\right)(X_{([np]+1)} - X_{([np])})$$

其中 $[x]$ 表示不大于 x 的最大整数。

本节介绍一些有关顺序统计量的基本知识。

1.4.1 顺序统计量的精确分布

本节始终考虑独立同分布 (iid) 的样本 $X_1, \cdots, X_n$。假设它们的总体分布函数为 $F(x)$，而其顺序统计量 $X_{(r)}$ 的分布函数为 $F_r(x)$，分布密度函数为 $f_r(x)$。我们有

$$F_r(x) = P(X_{(r)} < x)$$

$$=\sum_{j=r}^{n}\binom{n}{j}(F(x))^j(1-F(x))^{n-j}$$

$$=\frac{n!}{(r-1)!(n-r)!}\int_0^{F(x)}t^{r-1}(1-t)^{n-r}\mathrm{d}t$$

上面最后一个等式被一些书作为 $F_r(x)$ 的表达式. 如记 $X_{(r)}$ 的分布密度函数为 $f_r(x)$，则

$$f_r(x)=\frac{n!}{(r-1)!(n-r)!}(F(x))^{r-1}(1-F(x))^{n-r}f(x)$$

在 $r=n,r=1$ 的特别情况，我们有极大值和极小值的分布函数和密度函数

$$F_n(x)=(F(x))^n,\qquad f_n(x)=n(F(x))^{n-1}f(x)$$

$$F_1(x)=1-(1-F(x))^n,\qquad f_1(x)=n(1-F(x))^{n-1}f(x)$$

在 $F(x)$ 已知时，可用二项分布表或 Beta 不完全积分表来求 $F_r(x)$.

下面我们来求两个统计量 $X_{(r)}$ 和 $X_{(s)}$ 的联合分布 $F_{r,s}(x,y)\quad(r<s)$. 当 $x\leqslant y$ 时，

$$F_{r,s}(x,y)=P(X_{(r)}\leqslant x,\quad X_{(s)}\leqslant y)$$

$$=\sum_{l=s}^{n}\sum_{k=r}^{l}\frac{n!}{k!(l-k)!(n-l)!}(F(x))^k\cdot(F(y)-F(x))^{l-k}(1-F(y))^{n-l}$$

当 $x>y$ 时，

$$F_{r,s}(x,y)=P(X_{(r)}\leqslant x,\quad X_{(s)}\leqslant y)=P(X_{(s)}\leqslant y)=F_s(y)$$

即

$$F_{r,s}(x,y)=\begin{cases}\displaystyle\sum_{l=s}^{n}\sum_{k=r}^{l}\frac{n!}{k!(l-k)!(n-l)!}\\ \cdot(F(x))^k(F(y)-F(x))^{l-k}(1-F(y))^{n-l}, & x\leqslant y\\ F_s(y), & x>y\end{cases}$$

如相应的密度存在，则有

$$f_{r,s}(x,y)=\begin{cases}\displaystyle\frac{n!}{(r-1)!(s-r-1)!(n-s)!}\\ \cdot(F(x))^{r-1}(F(y)-F(x))^{s-r-1}(1-F(y))^{n-s}f(x)f(y), & x\leqslant y\\ 0, & x>y\end{cases}$$

特别，当 $r=1,s=n$ 时，

$$f_{1,n}(x,y)=\begin{cases}n(n-1)(F(y)-F(x))^{n-2}f(x)f(y), & x\leqslant y\\ 0, & x>y\end{cases}$$

对于两个以上顺序统计量的分布，因不常用，我们仅给出所有 n 个统计量的联合密度：

$$f_{1,2,\cdots,n}(x_1,\cdots,x_n)=\begin{cases}n!\displaystyle\prod_{i=1}^{n}f(x_i), & x_1<\cdots<x_n\\ 0, & \text{否则}\end{cases}$$

在随机摸拟中，产生某一分布的随机数是关键. 实际上该随机数是通过(0,1)上的均匀分布的随机数变换而得，而后者可由计算机的标准程序而得. 其理论依据为：

定理 1.1 如随机变量 X 具有连续分布函数 $F(x)$，则 $Y=F(X)$ 有(0,1)上的均匀分布 $U(0,1)$.

该结论的证明留给读者.

由此，如$X_{(1)},\cdots,X_{(n)}$为来自连续分布$F(x)$的顺序统计量，则$F(X_{(1)})\leqslant\cdots\leqslant F(X_{(n)})$为来自$U(0,1)$的顺序统计量. 注意，此结论的逆也对，即如果$F(x)$连续，$U\sim U(0,1)$，则$F^{-1}(U)\sim F(x)$. 随机模拟中有一种方法就利用这一性质.

1.4.2 顺序统计量的极限分布

显然，样本均值，p分位数(包括前面讲的"五数")等都是顺序统计量的线性组合. 我们因此考虑一般的统计量

$$T_n=\sum_{i=1}^{n}C_{ni}X_{(i)}$$

的分布. 因为其精确分布可由一变换求得，故这里只给出极限分布的结果. 令$J(u)$是一定义在(0,1)上的实函数. 权C_{ni}为

$$C_{ni}=\frac{1}{n}J\left(\frac{i}{n+1}\right),\qquad i=1,\cdots,n$$

我们有

定理 1.2 如果

1. $F(x)$在R上处处连续；
2. $\int_{-\infty}^{+\infty}|x|\mathrm{d}F(x)<\infty$；
3. $J(u)$在[0,1]上除有限个第一类间断点外，处处连续；
4. 除有限个例外点，$J'(u)$在[0,1]上处处连续，如在例外点上令$J'(u)=0$，则$J'(u)$在[0,1]上为有界变差；
5. 记$G(x)=F^{-1}(x)\equiv\inf\{y;F(y)\geqslant x\}$，

$$\sigma^2\equiv\int_0^1\int_0^1|J(s)J(t)|\min(s,t)(1-\max(s,t))\mathrm{d}G(s)\mathrm{d}G(t)<\infty$$

则

$$\sqrt{n}\left(T_n-\int_{-\infty}^{+\infty}xJ(F(x))\mathrm{d}F(x)\right)\xrightarrow{\mathscr{L}}N(0,\sigma^2)$$

证明请见[1].

由此定理可证样本均值的极限分布为正态，这与中心极限定理的结论是一致的.

1.4.3 顺序统计量的充分完全性

我们关于总体分布F的知识，全部来源于样本$X_1,\cdots,X_n$. 而统计量又是样本中信息的浓缩和概括. 如果统计量能保留所有样本中关于F的信息，则该统计量称为充分的(确切的定义参见§1.2.1). 这个概念是Fisher于1925年提出的. Neyman和Halmos的因子分解定理可用来验证一个统计量是否充分. 关于顺序统计量的充分性我们有如下的定理.

定理 1.3 对于分布族$\mathscr{F}$，$\forall F\in\mathscr{F}$，设$X_1,\cdots,X_n$为来自F的样本，只要$X_1,\cdots,X_n$是独立同分布的，则不论$\mathscr{F}$如何，$X_{(1)},\cdots,X_{(n)}$关于$\mathscr{F}$都是充分的.

完全性主要用于检验无偏估计的唯一性. 下面的定理给出了顺序统计量为完全的充要条件.

定理 1.4 假定$X_1,\cdots,X_n$为来自分布族$\mathscr{F}$的分布为F的独立同分布样本. 如果

(1) $\mathscr{F}$是凸的，即对于任意该族中的分布F_1和F_2及$0\leqslant\lambda\leqslant1$，有$\lambda F_1+(1-\lambda)F_2\in\mathscr{F}$；

(2) 对任意的 $a<b, S=[a,b)$，由 $F(b)-F(a)>0$ 可导出 $P(X_1<x|X_1\in S)\in\mathscr{F}$，则该样本的顺序统计量是关于 $\mathscr{F}$ 完全的.

上面两定理的证明见[8].

由定理 1.4 不难验证，许多统计分布族的顺序统计量是完全的.

1.4.4 极值统计量的分布

极值统计量在实际中有许多应用，比如人们关心河流的最高或最低水位，因为这与防汛和航运有紧密联系，气候和地质的极端条件对工程设计有重大影响等等. 极值分布是统计的一个分支，我们在此仅介绍一下极值统计量分布的三种类型. 称两个分布 F_1 和 F_2 是同类的，如果存在常数 $a>0$ 及 b，使得对所有的 x，有 $F_1(x)=F_2(ax+b)$. 这是一个等价关系，因为它满足自反性、对称性和传递性. 下面我们不加证明地引入 Gnedenko(1943) 的关于极值统计量分布函数的分类定理.

定理 1.5 如 G 为一连续函数的极大值分布，则它必属下面三类型之一，这里 $\alpha>0$，是个常数.

[*I*] $$G_1(x)=\exp\{-e^{-x}\},\quad -\infty<x<\infty$$

[*II*] $$G_2(x)=\begin{cases}\exp(-x^{-\alpha}), & x>0\\ 0, & x\leqslant 0\end{cases}$$

[*III*] $$G_3(x)=\begin{cases}\exp(-(-x)^{\alpha}), & x\leqslant 0\\ 1, & x>0\end{cases}$$

连续函数的极小值分布有类似分类，读者可自己得出. 其中最常见的是 *I* 型分布. 极值统计主要研究一个分布属于某个类型时需满足什么条件及如何估计参数等等.

1.5 U 统计量的基本知识

1.5.1 单样本 U 统计量的定义

设 $X_1,\cdots,X_n$ 为来自 $F(x)$ 的独立同分布样本. 假定我们对 F 的某参数 $\theta(F)$ 感兴趣，希望找到作为顺序统计量函数的 $\theta(F)$ 的一个无偏估计量，这就导致 U 统计量的引进. 有的 U 统计量是独立于总体分布的，有的可近似地看成非参数的(大样本时).

假定统计量 $h^*(X_1,\cdots,X_m)$ 是 $\theta(F)$ 的无偏估计，我们称 $h^*(\cdot)$ 为 $\theta(F)$ 的核. 我们总可假设核为对称的，即对任意的 $1,\cdots,m$ 的排列 $(\alpha_1,\cdots,\alpha_m)$，

$$h^*(x_1,\cdots,x_m)=h^*(x_{\alpha_1},\cdots,x_{\alpha_m})$$

这是因为总可以构造对称的核

$$h(x_1,\cdots,x_m)=\frac{1}{m!}\sum^{*}h^*(x_{\alpha_1},\cdots,x_{\alpha_m})$$

这里 $\sum^{*}$ 是对所有 $1,\cdots,m$ 的排列 $(\alpha_1,\cdots,\alpha_m)$ 求和. 对于对称的核 h，U 统计量定义为

$$U_n = \frac{1}{\binom{n}{m}} \sum_{1 \leqslant \alpha_1 < \cdots < \alpha_m \leqslant n}^{**} h(X_{\alpha_1}, \cdots, X_{\alpha_m})$$

其中 $\sum^{**}$ 表示对所有的从$(1,\cdots,n)$中取出的 m 个数的排列$(\alpha_1,\cdots,\alpha_m)$且满足$\alpha_1 < \cdots < \alpha_m$ 的 $\alpha_1,\cdots,\alpha_m$ 求和.

两个最简单的 U 统计量的例子就是样本均值和样本方差.它们分别由取 $h(y) = y, m = 1$ 及 $h(y_1,y_2) = \frac{(y_1 - y_2)^2}{2}$ 所得,并且,在定理 1.4 的条件下,它们是一致最小方差无偏估计量(UMVUE).实际上,如顺序统计量完全,则相应的 U 统计量是 UMVUE,而且是唯一的.从上面 U 统计量的构造,可知如何由一个无偏估计量去产生 UMVUE.

1.5.2 两样本 U 统计量的定义

前面介绍了单样本的 U 统计量.当我们处理两样本时,可类似定义两样本的 U 统计量.

设 $X_1,\cdots,X_m$ 和 $Y_1,\cdots,Y_n$ 为分别源自 $F(x)$ 和 $G(y)$ 的独立随机样本.假定统计量 $h^*(X_1,\cdots,X_r,Y_1,\cdots,Y_s)$ 是 $\theta(F,G)$ 的无偏估计,我们称 $h^*(\cdot)$ 为 $\theta(F,G)$ 的核.同样,我们可假设核为对称的.而两样本 U 统计量定义为

$$U(X_1,\cdots,X_m,Y_1,\cdots,Y_n) = \frac{1}{\binom{m}{r}\binom{n}{s}} \sum_{\alpha}^{**} \sum_{\beta}^{**} h(X_{\alpha_1},\cdots,X_{\alpha_r},Y_{\beta_1},\cdots,Y_{\beta_s})$$

这里,符号 $\sum^{**}$ 和前面一样,只不过分别是对 α 和 β 而已.下面给几个例子.

例 1.5 设 $X_1,\cdots,X_n$ 为来自 F 的独立同分布样本,如取 $m = 1$,核函数 $h(X_1) = I(X_1 > 0)$,则单样本 U 统计量 U_n 为

$$U_n = \frac{1}{n}\sum_{i=1}^{n} I(X_i > 0) \equiv \frac{1}{n} S$$

其中统计量 S 描述了样本中大于 0 的个数,我们称之为符号统计量(sign statistic),这是非参数统计中经常用到的统计量之一.

例 1.6 设 $X_1,\cdots,X_n$ 为来自 $F(x-\theta)$ 的随机样本,$F(\cdot)$ 处处连续且关于原点对称.如取 $m = 2, h(X_1,X_2) = I(X_1 + X_2 > 0)$,则以 $h(X_1,X_2)$ 为核的单样本 U 统计量 U_m 为

$$U_n = \frac{1}{\binom{n}{2}} \sum_{i<j} I(X_i + X_j > 0)$$

如记

$$R_i^+ = \sum_{j=1}^{n} I(|X_j| \leqslant |X_i|), \qquad r = \sum_{i=1}^{n} I(X_i > 0)$$

则通过一定的推理计算可得

$$U_n = \frac{1}{\binom{n}{2}} \left(\sum_{i=1}^{n} R_i^+ - r\right)$$

通过以后的学习我们知道,这也是一个很著名的非参数检验统计量——Wilcoxon 符号秩统计量.

例 1.7 假定 $X_1,\cdots,X_m \overset{iid}{\sim} F$ 和 $Y_1,\cdots,Y_n \overset{iid}{\sim} G$ 为二独立样本. 则以 $h(x,y)=I(y>x)$ 为核的两样本 U 统计量为

$$U_{m,n}=\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}h(Y_j>X_i)=\frac{1}{mn}\sum_{i,j}I(Y_j-X_i>0)=\frac{W}{mn}$$

其中 $W=\sum_{i,j}I(Y_j-X_i>0)$ 称为 Mann-Whitney 统计量. 通常, 我们感兴趣的是检验 H_0: $F=G$, 把两样本合起来形成一 $N\equiv m+n$ 个数据的样本, 在零假设下, 它们是独立同分布 (iid) 的. 把它们按次序排成 $Z_1<\cdots<Z_N$. 对 $j=1,\cdots,n$ 记

$$R_j=\sum_{i=1}^{N}I(Z_i\leqslant Y_j)$$

则

$$W=\sum_{i=1}^{m}\sum_{j=1}^{n}I(Y_j>X_i)=\sum_{j=1}^{n}R_j-\frac{n(n+1)}{2}$$

这个统计量以后还要用到, 它说明Mann-Whitney统计量只与 Y 样本在合样本中的位置有关.

1.6 渐近相对效率

在假设检验中, 当显著性水平固定时, 比较好的检验有较大的势. 然而, 势的大小依赖于样本的大小. 人们自然会认为, 当势相等时, 样本小的效率高. 类似地, 在比较估计量时, 通常认为方差小的好. 但方差也依赖于样本大小. 人们同样会认为, 当估计量的方差相等时, 样本小的效率高. 为获得同样的势或同样的方差, 两个检验统计量或两个估计量所用的样本数目的比值就是相对效率. 这个比值的极限为渐近相对效率(asymptotic relative efficiency —ARE). 虽然这里说的是样本数目的比, 但实际上, 如使这两个有关的样本数保持相等, 势或方差的比也导致同样的相对效率的概念.

我们用检验来引入 ARE 的概念. 如检验 $H_0:\theta=0\leftrightarrow H_1:\theta>0$, 令 $V_n^{(1)},V_n^{(2)}$ 表示两个检验统计量. 假定拒绝区域为($V_n^{(i)}\geqslant k_n^{(i)}$), $i=1,2$. 如果在 H_0 下, 当 $n\to\infty$ 时, $P(V_n^{(i)}\geqslant k_n^{(i)})\to\alpha$, 我们说该检验有渐近水平 α. 对固定的 $\beta(\alpha<\beta<1)$, 如 $\{\theta_j^{(i)}\}$ 为一个备选假设序列, 而且随着序列 $\{n_j^{(i)}\}$, $i=1,2$ 趋于无穷, $\theta_j\to 0$, 使得对 $i=1,2$ 有 $P(V_n^{(i)}\geqslant k_n^{(i)})\to\beta$. 如果极限

$$e_{12}\equiv\lim_{j\to\infty}\frac{n_j^{(2)}}{n_j^{(1)}}$$

存在而且独立于 $\{\theta_j\}$, α,β, 则称 e_{12} 为 $V_n^{(1)}$ 相对于 $V_n^{(2)}$ 的渐近相对效率. 简记为 ARE($V_n^{(1)},V_n^{(2)},F$). 这是Pitman 于 1948 年提出的, 故又称 Pitman 效率. 这个思想简单, 但不好用. 下面介绍一个实用的求 e_{12} 的步骤.

假定下列五个条件(Pitman 条件)成立:

(1) V_n 是一个相容检验统计量, 即当 $n\to\infty$ 时, 对 $\theta\in\Theta_1$, 势函数 $\beta(\theta,V_n)\to 1$.

(2) 存在序列 $\{\mu_n(\theta)\}$ 和 $\{\sigma_n(\theta)\}$, 使得对在 $\theta=0$ 的一个邻域中一致渐近地有

$$\frac{V_n-\mu_n(\theta)}{\sigma_n(\theta)}\sim N(0,1)$$

(3) 存在导数 $\mu'_n(0)=\left.\frac{\mathrm{d}\mu_n(\theta)}{\mathrm{d}\theta}\right|_{\theta=0}$.

(4) 对于趋于零的序列$\{\theta_n\}$,当 $n\to\infty$ 时

$$\frac{\sigma_n(\theta_n)}{\sigma_n(0)}\to 1,\qquad \frac{\mu'_n(\theta_n)}{\mu'_n(0)}\to 1$$

(5)

$$\frac{\mu'_n(0)}{\sqrt{n}\,\sigma_n(0)}\to c>0$$

这里 c 称为 V_n 的效率(efficacy),c^2 称为效率因子.

前面所讲的 $V_n^{(1)}$相对于 $V_n^{(2)}$的渐近相对效率等于

$$e_{12}\equiv\lim_{j\to\infty}\frac{n_j^{(2)}}{n_j^{(1)}}=\frac{c_1^2}{c_2^2}$$

这里 c_i 为 $V_n^{(i)}$, $i=1,2$ 的效率. 因为求效率所需的 $\mu'_n(0)$ 和 $\sigma_n(0)$ 都不难,故渐近相对效率也可得到.

以后我们将会看到,利用渐近相对效率这个度量,可以看出非参数统计方法在许多情况下有着不可比拟的优越性.

1.7 阅读知识

1.7.1 顺序统计量

通过定理 1.3、1.4 可以知道,顺序统计量在许多情况下,尤其是在非参数统计模型中,都是充分完全统计量,故它在非参数统计中占有极重要的地位. 有关其详细的知识可见[8]. 下面,我们仅给出一类特殊的顺序统计量的线性组合——样本分位数的极限分布.

以下设 $X_1,\cdots,X_n$ 为来自某分布函数 $F(x)$ 的样本,$f(x)$ 为概率密度. 以 $X_{(1)},\cdots,X_{(n)}$ 表示顺序统计量. 从样本 p 分位数 m_p 的表达式可以看出,它是 $X_{([np])}$ 与 $X_{([np]+1)}$ 的线性组合. 当 n 很大时,可认为 $m_p\doteq X_{([np]+1)}\equiv\hat{m}_{n,p}$ 下面我们考虑 $\hat{m}_{n,p}$ 的极限性质.

定理 1.6 (样本分位数的 Bahadur 表示) 以 F_n 记 $X_1,\cdots,X_n$ 的经验分布,ξ_p 表示总体的 p 分位数,如果 $f(\xi_p)>0$ 且 $f(\cdot)$ 在 ξ_p 点连续,则当 $n\to\infty$ 时,

$$\sqrt{n}\left\{\hat{m}_{n,p}-\xi_p+\frac{F_n(\xi_p)-p}{f(\xi_p)}\right\}\xrightarrow{P}0$$

证明 见[1].

从上一定理可以看出,样本 p 分位数可以用

$$\xi_p-\frac{F_n(\xi_p)-p}{f(\xi_p)}$$

来近似表示,它说明了样本 p 分位数与总体 p 分位数之间的差距. 关于样本分位数还有如下的极限分布.

定理 1.7 设 ξ_p 为总体的 p 分位数，$f(\xi_p) > 0$，且 $f(x)$ 在 ξ_p 点连续，则当 $n \to \infty$ 时，有

$$\sqrt{n}(\hat{m}_{n,p} - \xi_p) \xrightarrow{\mathscr{L}} N\left(0, \frac{p(1-p)}{f^2(\xi_p)}\right)$$

特别当 $p = 0.5$ 时，

$$\sqrt{n}(X_{\text{med}} - \xi_{0.5}) \xrightarrow{\mathscr{L}} N\left(0, \frac{f^{-2}(\xi_{0.5})}{4}\right)$$

证明 见[8].

上一极限分布在统计推断中是非常有用的一个结论.

1.7.2 U 统计量

在极限理论中我们知道，随机变量的独立同分布和的极限分布是正态分布. 而从 U 统计量的形式上看，U 统计量也是独立同分布和的一种推广. 近代研究表明，这种看法是正确的. 下面我们则不加证明地引进有关 U 统计量的大样本性质.

下设 U_n 表示以 $h(X_1,\cdots,X_m)$ 为对称核，基于来自分布 $F(x)$ 的独立同分布样本 $X_1,\cdots,X_n$ 的单样本 U 统计量，记

$$\theta(F) = E_F h(X_1,\cdots,X_m)$$

由 U 统计量的定义，显然有 $E_F U_n = \theta(F)$.

为求 U_n 的方差，任给 $1 \leqslant k \leqslant m$，定义

$$\begin{aligned} h_k(x_1,\cdots,x_k) &= E_F(h(X_1,\cdots,X_m) \mid X_1 = x_1,\cdots,X_k = x_k) \\ &= E_F h(x_1,\cdots,x_k,X_{k+1},\cdots,X_m) \end{aligned}$$

记

$$\sigma_k^2 = \text{Var} h_k(X_1,\cdots,X_k), \quad k = 1,\cdots,m$$

可以证明，如果 $E_F h^2(X_1,\cdots,X_m) < \infty$，则 $\sigma_k^2 < \infty$. 不妨设 $\theta(F) = 0$，则

$$\begin{aligned} &\text{Var}\left(\binom{n}{m} U_n\right) = E\left(\binom{n}{m} U_n\right)^2 \\ &= E\left(\sum_{1 \leqslant i_1 < \cdots < i_m \leqslant n} h(X_{i_1},\cdots,X_{i_m})\right)^2 \\ &= \sum^{*} Eh(X_{i_1},\cdots,X_{i_m}) Eh(X_{j_1},\cdots,X_{j_m}) + \sum^{**} Eh(X_{i_1},\cdots,X_{i_m}) h(X_{j_1},\cdots,X_{j_m}) \\ &= \sum_{k=1}^{m} \sum^{***} Eh(X_{i_1},\cdots,X_{i_m}) h(X_{j_1},\cdots,X_{j_m}) \end{aligned}$$

其中求和 $*$，$**$，$***$ 分别表示在如下集合中进行：

$$S_1 = \{1 \leqslant i_1 < i_2 < \cdots < i_m \leqslant n, \quad 1 \leqslant j_1 < j_2 < \cdots < j_m \leqslant n; \quad \forall\, k,l, \quad i_k \neq j_l\}$$

$$S_2 = \{1 \leqslant i_1 < i_2 < \cdots < i_m \leqslant n, \quad 1 \leqslant j_1 < j_2 < \cdots < j_m \leqslant n; \quad \exists\, k,l, \quad i_k = j_l\}$$

$S_3 = \{$在 $1 \leqslant i_1 < i_2 < \cdots < i_m \leqslant n$ 与 $1 \leqslant j_1 < j_2 < \cdots < j_m \leqslant n$ 中恰有 k 个相同$\}$

又因为

$$\begin{aligned} &E(h(X_1,\cdots,X_k,X_{k+1},\cdots,X_m) h(X_1,\cdots,X_k,X_{m+1},\cdots,X_{2m-k})) \\ &= E(E(h(X_1,\cdots,X_k,X_{k+1},\cdots,X_m) h(X_1,\cdots,X_k,X_{m+1},\cdots,X_{2m-k})) \mid X_1,\cdots,X_k) \\ &= Eh_k^2(X_1,\cdots,X_k) = \sigma_k^2 \end{aligned}$$

所以

$$\mathrm{Var}U_n = \frac{1}{\binom{n}{m}}\sum_{k=1}^{m}\binom{m}{k}\binom{n-m}{m-k}\sigma_k^2$$

对于两样本 U 统计量,如定义

$$\begin{aligned}&h_{k,l}(x_1,\cdots,x_k;y_1,\cdots,y_l)\\&=E(h(X_1,\cdots,X_r;Y_1,\cdots,Y_s)\mid X_1=x_1,\cdots,X_k=x_k;Y_1=y_1,\cdots,Y_l=y_l)\\&\sigma_{k,l}^2\equiv \mathrm{Var}h_{k,l}(X_1,\cdots,X_k;Y_1,\cdots,Y_l)\end{aligned}$$

则

$$\mathrm{Var}U_{m,n} = \frac{\sum_{k=0}^{r}\sum_{l=0}^{s}\binom{r}{k}\binom{m-r}{r-k}\binom{s}{l}\binom{n-s}{s-l}\sigma_{k,l}^2}{\binom{m}{r}\binom{n}{s}}$$

利用组合数的变换可以证明,当 $n\to\infty$ 时

$$\mathrm{Var}U_n = \frac{m^2}{n}\sigma_1^2 + O(n^{-2})$$

当 $m\to\infty, n\to\infty$ 时

$$\mathrm{Var}U_{m,n} = \frac{r^2}{m}\sigma_{1,0}^2 + \frac{s^2}{n}\sigma_{0,1}^2 + O\left(\frac{1}{m^2}+\frac{1}{n^2}\right)$$

有了上面的数字特征之后,Hoeffding(1948)利用独立同分布之和去逼近 U 统计量这一方法,证明了下面的关于 U 统计量的渐近性质.

定理 1.8 如果 $Eh^2(X_1,\cdots,X_m)<\infty$,且 $\sigma_1^2>0$,则当 $n\to\infty$ 时,有

$$\sqrt{n}\,(U_n-\theta(F)) \xrightarrow{\mathscr{L}} N(0,m^2\sigma_1^2)$$

证明 见[1].

对于两样本 U 统计量有下面类似的结论.

定理 1.9. 如果 $Eh^2(X_1,\cdots,X_r;Y_1,\cdots,Y_s)<\infty, \sigma_{1,0}^2>0, \sigma_{0,1}^2>0$,记

$$N=m+n,\qquad \sigma_{m,n}^2\equiv N\left(\frac{r^2}{m}\sigma_{1,0}^2+\frac{s^2}{n}\sigma_{0,1}^2\right)$$

则当 $m\to\infty, n\to\infty$ 时,有

$$\frac{\sqrt{N}(U_{m,n}-\theta(F,G))}{\sigma_{m,n}} \xrightarrow{\mathscr{L}} N(0,1)$$

证明 见[1].

有了上面的极限分布,则我们可以利用 U 统计量进行假设检验和估计,感兴趣的读者可参见[1].

1.8 习　　题

1. 任给 $0\leqslant t\leqslant 1$,证明:$\forall\ 0\leqslant r\leqslant n$,

$$\sum_{j=r}^{n}\binom{n}{j}t^j(1-t)^{n-j}=\frac{n!}{(r-1)!(n-r)!}\int_0^t x^{r-1}(1-x)^{n-r}\mathrm{d}x$$

2. 设随机变量 $\xi\sim F(x)$，试证明，如 $F(x)$ 并非处处连续，则 $F(\xi)$ 不服从 $(0,1)$ 上的均匀分布。

3. 当总体为 $(0,1)$ 上的均匀分布，且 n 为偶数时，求出样本中位数的概率密度。

4. 举一个简单例子证明，样本中位数不一定是总体中位数的无偏估计。

5. 设 $U_1\leqslant U_2\leqslant\cdots\leqslant U_n$ 为来自 $(0,1)$ 上均匀分布的顺序样本，证明：$\forall\ 1\leqslant r<s\leqslant n$，$U_s-U_r$ 与 U_{s-r} 同分布。

6. 设总体分布族 $\mathscr{F}=\{u(\theta_1,\theta_2):-\infty<\theta_1<\theta_2<+\infty\}$，证明，当 $n=2$ 时，$(X_{(1)},X_{(2)})$ 为完全统计量，而当 $n\geqslant 3$ 时，$(X_{(1)},\cdots,X_{(n)})$ 却不是完全统计量（这说明在参数统计中，顺序统计量不一定是充分完全统计量）。

7. 设分布函数 $F(x)$ 连续，$X_1,\cdots,X_n$ 为来自 $F(x)$ 的样本，试证明

(1) $-2\sum\limits_{i=1}^{n}\ln F(X_i)\sim\chi^2(2n)$；

(2) $F(X_{(k)})\sim B_{\text{I}}(k,n-k+1)$，

其中 B_{I} 表示 I 型 Bate 分布。

8. 设 $F(x)$ 为具有连续单峰概率密度函数 $f(x)$ 的关于 θ 对称的分布函数。请证明以下两个结论：

(1) 对于 $0<\alpha<\frac{1}{2}$，设 $k=[n\alpha]$，则 α- 切尾均值

$$V(X_1,\cdots,X_n)=\frac{1}{n-2k}\sum_{i=k+1}^{n-k}X_{(i)}$$

有如下的极限分布：

$$\frac{\sqrt{n}\,(V(X_1,\cdots,X_n)-\theta)}{\sigma}\xrightarrow{\mathscr{L}}N(0,1)$$

其中

$$\sigma^2=\frac{\left(\int_{-\xi_{1-\alpha}}^{\xi_{1-\alpha}}x^2\mathrm{d}F(x)+\alpha\xi_{1-\alpha}^2\right)}{(1-2\alpha)^2},\qquad \xi_{1-\alpha}=\sup\{y:F(y)<1-\alpha\}$$

(2) 对于 $\alpha\in(0,\frac{1}{2})$，设 $k=[n\alpha]$，则 α-Winsor 化均值

$$W(X_1,\cdots,X_n)=\frac{1}{n}\left(\sum_{i=k+1}^{n-k}X_{(i)}+kX_{(k+1)}+kX_{(n-k)}\right)$$

有如下的极限分布：

$$\frac{\sqrt{n}\,(W(X_1,\cdots,X_n)-\theta)}{\sigma}\xrightarrow{\mathscr{L}}N(0,1)$$

其中

$$\sigma^2=\int_{-\xi_{1-\alpha}}^{\xi_{1-\alpha}}x^2\mathrm{d}F(x)+2\alpha\left(\xi_{1-\alpha}-\frac{\alpha}{f(\xi_{1-\alpha})}\right)^2$$

9. 试证明例 1.6 中的 U 统计量的等价表达式。

10. 试证明例 1.7 中的统计量 W 的等价表达式。

11. 设 $X_{(1)}\leqslant\cdots\leqslant X_{(n)}$ 为来自 $(0,1)$ 上均匀分布的顺序统计量，求 $\mathrm{Cov}(X_{(r)},X_{(s)})$，$\forall\ 1\leqslant r\leqslant s\leqslant n$。

12. 设 (X,Y) 具有二元连续分布，X,Y 的边际分布分别为 $F(x)$ 和 $G(y)$，试证明：$(F(X),G(Y))$ 在 $(0,1)\times(0,1)$ 上均匀分布的充要条件为 X 与 Y 独立。

13. 设 $X_{(1)}\leqslant\cdots\leqslant X_{(n)}$ 为来自概率密度函数为

$$f(x)=\frac{\mathrm{e}^{-x}}{(1+\mathrm{e}^{-x})^2},\quad x\in(-\infty,+\infty)$$

的 logistic 分布的样本，试求其中位数的概率密度.

14. 设 $X_{(1)} \leqslant \cdots \leqslant X_{(n)}$ 为来自具有概率密度

$$P(X_i = t) = \begin{cases} (1-p_i)^{t-1}p_i, & t = 1,2,\cdots \\ 0, & \text{其它} \end{cases}$$

的几何分布的 iid 样本，试求 $X_{(1)}$ 的分布.

15. 设 $X_{(1)} \leqslant \cdots \leqslant X_{(n)}$ 为来自某连续分布 $F(x)$ 的 iid 样本，且具有概率密度函数 $f(x)$，如定义

$$U_i = \frac{F(X_{(i)})}{F(X_{(i+1)})}, \quad i = 1,\cdots,n-1, \quad U_n = F(X_{(n)})$$

则证明 $U_1, U_2^2, \cdots, U_n^n$ 为来自 $(0,1)$ 上均匀分布的 iid 样本.

16. 设 $X_{(1)} \leqslant \cdots \leqslant X_{(n)}$ 为来自某连续分布 $F(x)$ 的次序样本，令 $\tau_{(i)} = EX_{(i)}$，$i = 1,\cdots,n$（假设期望存在），又设 $Y_{(1)} \leqslant \cdots \leqslant Y_{(n)}$ 为来自 $F\left(\frac{x-\mu}{\sigma}\right)$ 的顺序样本（$\mu > 0, \sigma > 0$ 为常数），试证：$EY_{(i)} = \mu + \sigma\tau_{(i)}$，$i = 1,\cdots,n$.

17. 设 $X_1 \leqslant \cdots \leqslant X_n$ 为来自某分布函数 $F(x)$ 的 iid 样本，对于下面的参数 θ，试求基于 $X_1,\cdots,X_n$ 的关于 θ 的 U 统计量

(1) $P(|X_1| > 1)$；

(2) $P(X_1 + X_2 + X_3 > 0)$；

(3) $E(X_1X_2)^3$；

(4) $E(X_1 - X_2)^4$；

(5) $\mathrm{Cov}(X_1, X_2)$.

18. 设 $X_1 \leqslant \cdots \leqslant X_n$ 为来自某分布函数 $F(x)$ 的 iid 样本，对于参数 $\theta = P(X_1 + X_2 > 0)$，我们可以取例 1.6 中的核函数，但也可以取如下的核函数：

$$h_1(x) = 1 - F(-x)$$

试证之，并说明 $h_1(x)$ 是否为对称核？

19. 设 $X_1 \leqslant \cdots \leqslant X_m$ 和 $Y_1 \leqslant \cdots \leqslant Y_n$ 为分别来自连续分布 $F(x)$ 和 $G(y)$ 的相互独立的 iid 样本，$\theta = P(X_1 + X_2 < Y_1 + Y_2)$.

(1) 证明在 $H_0: F = G$ 之下，$\theta = \frac{1}{2}$；

(2) 试求关于 θ 的 U 统计量.

20. 设 $X_1 \leqslant \cdots \leqslant X_n$ 和 $Y_1 \leqslant \cdots \leqslant Y_n$ 为分别来自连续分布的相互独立的样本，试求 $\theta = \mathrm{Var}(X) + \mathrm{Var}(Y)$ 的 U 统计量.

21. 利用表 1.1 中的数据，

(1) 构造一班的茎叶图，并写出其层；

(2) 由(1)中的茎叶图求出一班成绩的中位数；

(3) 计算一班的平均分；

(4) 计算一班的下上四分位数 Q_l 及 Q_u；

(5) 检验一班成绩是否有异常值；

(6) 试画出一班成绩的盒子图.

22. 利用表 1.1 的数据，给出一班的盒子图.

23. 在同一个图中给出一、二、三班的盒子图.

24. 一个超级商场的经理对顾客在商场中的逗留时间感兴趣. 现随机地测量 20 人的逗留时间为（单位：分钟）：

34 28 32 24 38 16 8 24 50 26

12 20 22 42 30 26 32 28 2 26

(1) 试写出这些数据的茎叶图及层；

(2) 求这些顾客逗留时间的五数概括；

(3) 检测这些数据是否有异常点；

(4) 画出盒子图；

(5) 画出直方图.

25. 某大学新生入学的第一天，有人询问了若干名新生衣服口袋中带有多少钱，现记录如下(单位：元)：

男生：	81	26	8	10	0	20	14	33	50	10
	0	12	23	55	28	56	53	55	2	
女生：	4	128	1	0	73	2	8	3	24	94
	30	39	10	146	0	37	10	22	6	8
	10	47	33	7						

(1) 试画出男生与女生的背靠背茎叶图；

(2) 试分别求出男生与女生所带钱的五数概括；

(3) 试写出联合的茎叶图及五数概括；

(4) 检验联合数据是否有异常点.

26. 下面数据记录了美国在1986年的50个州及哥伦比亚特区的失业率(%)：

5.3	2.8	4.7	3.8	4.0	3.8	6.3	5.0	6.8	8.1
6.7	8.1	8.8	7.0	5.3	7.0	6.1	6.3	4.7	5.0
5.4	4.3	4.5	7.7	5.0	11.8	5.3	6.2	5.9	5.7
9.3	8.0	9.8	11.7	8.7	13.1	8.2	8.9	8.1	8.7
9.0	7.4	9.2	6.9	6.0	6.0	8.2	8.5	6.7	10.8
4.8									

试画出其盒子图.

27. 下面数据记录了美国在1976年的50个州及哥伦比亚特区的失业率(%)：

8.9	6.4	8.7	9.5	8.1	9.5	10.3	10.4	7.9	7.8
6.1	6.5	9.4	5.6	5.9	4.0	6.2	3.6	3.4	3.3
4.2	8.9	6.8	9.1	5.9	7.5	6.2	6.9	8.1	9.0
5.6	6.0	6.8	6.6	7.1	6.8	5.6	5.7	6.1	5.7
4.1	5.9	9.1	9.8	5.7	9.0	8.7	9.5	9.2	6.8
9.8									

试画出其盒子图，并与1986年失业率相比较，对这些差异也画出其盒子图.

28. 下面一组数据记录了在某地区工作的21名同志的月均收入(单位：元)：

819	779	575	665	481	599	493
454	392	534	479	296	345	244
349	279	361	438	194	301	189

(1) 试写出其茎叶图(注意:此时的茎叶图可以通过忽略其个位数字而得到,但不要四舍五入);

(2) 画出其盒子图.

29. 下面数据记录了某些人的身高数据(单位:厘米;且减去了100):

71.4	70.0	72.2	86.0	79.6	67.5	60.6
62.1	75.6	77.9	72.7	65.2	59.0	64.0
60.6	58.9	55.1				

试写出其茎叶图(注意:此时的茎叶图可以通过忽略其小数位而得到,但不要四舍五入).

第二章　单样本问题

2.1 引　　言

给定一组样本，最常见的统计问题就是对其总体分布的位置参数进行推断. 通常的位置参数是中位数或均值，故要对它进行假设检验、点估计或区间估计；有时数据可能与采集时的次序有关，故要发现数据的趋势；当对样本的随机性有怀疑时，又要检验其随机性等等. 另外在用传统的参数方法对位置参数进行推断时，人们假设总体是正态分布，或近似的正态分布，然后利用 t 检验或与其相关的点估计或区间估计. 但是关于总体是正态的假设并不一定合理，在小样本时，近似也不一定合适. 这时，如果用 t 检验，就可能会犯错误. 事实上，这是个很常见的错误. 对于数据的趋势或随机性等问题，不存在简单初等的参数方法. 但所有这些问题都有简单的非参数统计方法，它对总体分布并不作什么（或极少）假设，故有很大实用价值.

2.2 符 号 检 验

2.2.1 检验方法

符号检验是最简单、最古老的非参数方法. 我们先举一例子说明. 假定某地的 10 栋房屋出售价格（由低到高排列）为 56，69，85，87，90，94，96，113，118，179（单位：千美元），问该地区的平均房屋价格是否和人们相信的 8 万 4 千美元的水平大体一致. 用 M 来表示价格分布的中心（这里考虑中位数）. 如假设该分布是对称的，则 M 也是均值. 我们要检验 $H_0:M=M_0=84\leftrightarrow H_1:M\neq M_0=84$. 按照传统的参数方法，假设房屋价格 $X_1,\cdots,X_n(n=10)$ 为 iid 的 $N(M_0,\sigma^2)$ 分布，则 t 检验的检验统计量为 $T=\dfrac{\sqrt{n}(\overline{X}-M_0)}{S}$，$T$ 服从 $t(n-1)$ 分布. 这里 $\overline{X}$ 和 S^2 为样本均值和样本方差. 对于这组数据，$\bar{x}=99$，$s=33.186$，T 的值 $t=1.429$. 查表知 p 值为0.2，因此我们不拒绝零假设（对任何水平 $\alpha<0.2$）. 显然，无论答案是否合理，这种正态分布的假设是没有根据的. 我们现在作另一种考虑. 按照零假设，数据应以 $p=0.5$ 的概率位于中心 $M_0=84$ 的两边. 换句话说，样本中 X_i-M_0 符号为正的数目 S^+ 为二项分布 $\mathrm{B}(n,p)$. 同样地，样本中 X_i-M_0 符号为负的数目 S^- 也为二项分布 $\mathrm{B}(n,1-p)$. 对本例来说 S^- 和 S^- 的分布是一样的. 显然，当 S^+ 或 S^- 太大（或它们中小的一个太大）时，我们拒绝原假设. 我们原来的假设检验可等价地写成 $H_0:p_0=0.5\leftrightarrow H_1:p_0\neq0.5$. 这样问题就成为人们熟知的二项分布的检验问题. 因为它涉及符号，故称为符号检验(sign test). 令 $K=\min(S^-,S^-)$ 为我们的

检验统计量. 对本例 $K=2$,

$$P(K \leqslant 2 \mid n=10, p=0.5)=P(K=0)+P(K=1)-P(K=2)=0.0547$$

即 p 值为 $2 \times 0.0547=0.1094$. 因此我们不拒绝零假设(对任何水平 $\alpha<0.1094$). 对于单边检验 $H_0: M \leqslant M_0 \leftrightarrow H_1: M>M_0$, 显然, 当 S^+ 太小或 S^- 太大时拒绝零假设. 我们可用其中任何一个作为检验统计量(比如取 $K=S^+$). 在本例中, 我们如改用 $M_0=120$, 则 $S^+=1$. p 值为 $P(K=S^+ \leqslant 1)=0.011$. 我们因此可拒绝原假设(对任何水平 $\alpha>0.011$). 在实践中, 可能会遇到某些 $X_i=M_0$ 的情况. 这时仅需去掉这些值, 并相应地减少 n 的值.

2.2.2 大样本近似

当样本大的时候 (n 大时), 往往很难计算 p 值. 我们可用二项分布的正态近似, 即对于 $K \sim \mathrm{B}(n, p)$, 当 n 大时可近似地认为

$$Z=\frac{K-np}{\sqrt{np(1-p)}} \xrightarrow{\mathscr{L}} N(0,1)$$

在实用中, 要用连续改正量(continuity correction), 这是用连续分布来近似离散分布时所必需的. 对符号检验, $p=0.5$, 上面的 Z 值 z(相应于 K 的值 k)应为

$$z=\frac{k+C-0.5n}{0.5\sqrt{n}} \xrightarrow{\mathscr{L}} N(0,1)$$

这里 $C=0.5$ 或 $C=-0.5$ 依 $k<0.5n$ 或 $k>0.5n$ 而定. p 值为 $2P(Z \leqslant z)=2\Phi(z)$(对双边检验) 或 $P(Z \leqslant z)=\Phi(z)$(对单边检验 $H_0: M \leqslant M_0 \leftrightarrow H_1: M>M_0$). 这里 $\Phi(\cdot)$ 为标准正态分布函数, 其临界值可由附表 2 查得.

下面介绍另一例子. 某国 12 位总统的寿命(岁) 分别为 46, 57, 58, 60, 60, 63, 64, 67, 72, 78, 88, 90. 问该国总统寿命的中位数是否大于等于 $M_0=71.5$? (问题成为检验 $H_0: M \geqslant M_0 \leftrightarrow H_1: M<M_0$). 显然, 当 S^- 太小时拒绝原假设. 这里, $K=S^+=4$. 计算结果表明, 用二项分布算的 p 值为 0.1937, 而用正态近似算的 p 值为 0.1922. 结果类似.

2.2.3 基于符号检验的中位数的置信区间

借助于顺序统计量及层的概念, 加上二项分布 $\mathrm{B}(n, 0.5)$ 或其正态近似的概率计算, 很容易得到中位数 M 的置信区间. 假定我们有一样本以 $X_{(1)}, \cdots, X_{(n)}$ 为其顺序统计量. 最简单的置信区间是以其第一层为置信限的置信区间 $(X_{(1)}, X_{(n)})$. 相应的置信度为

$$P(X_{(1)} \leqslant M \leqslant X_{(n)})=1-P(M<X_{(1)})-P(M>X_{(n)})=1-\left(\frac{1}{2}\right)^{n-1}$$

对于上面总统寿命的例子, $n=12, x_{(1)}=46, x_{(12)}=90$; $1-0.5^{11}=0.9995$. 因而, 置信度为 0.9995 的置信区间为(46, 90).

我们知道, 人们总是希望置信区间小而同时置信度大, 但是不可能两全其美, 只能固定其一, 而使另一个尽可能地好. 上面例子的置信度虽大, 但区间也太大了. 为此我们可取不同层的数据作为置信限, 以满足事先给定的置信度 $1-\alpha$. 比如我们可用 $k+1$ 层的两个数据形成置信区间 $(X_{(k+1)}, X_{(n-k)})$, 同时满足 $P(K \leqslant k)+P(K>n-k) \leqslant \alpha$. 这个区间就是 M 的有 $[1-2P(K \leqslant k)] \times 100\%$ 置信度的置信区间. 对于上面总统寿命例子, 我们取不同的 k 值, 就得到

不同置信度的各种置信区间(见下表),这里的概率根据 Z 是二项分布而得.

k	$P(K=k)$	$P(K\leqslant k)$	$1-2P(K\leqslant k)$	置信区间
0	0.00024	0.00024	0.9995	(46,90)
1	0.0029	0.0031	0.9938	(57,88)
2	0.0161	0.0192	0.9616	(58,78)
3	0.537	0.0729	0.8542	(60,72)

在大样本时,我们可用正态分布近似二项分布:$\frac{2(K+1-0.5n)}{\sqrt{n}}\sim N(0,1)$. 因此,对给定的 α,取 $k+1\approx 0.5n-Z_{\frac{\alpha}{2}}\sqrt{\frac{n}{4}}$,这里 Z_{α} 满足 $1-\Phi(Z_{\alpha})=\alpha$,可查附表 2 得到. 对上例,如求置信度至少 95% 的置信区间,$\alpha=0.05$,取 $k+1\approx\frac{12}{2}-1.96\sqrt{\frac{12}{4}}\approx 2$. 注意,这里 $k+1$ 要取整数部分.

当然,置信限不一定非要取同一层的两个数. 这时,为了得到置信度为 $100(1-\alpha)\%$ 的置信区间 $(X_{(i)},X_{(j)})(i<j)$,我们可取 i,j 满足

$$1-\alpha=P(X_{(i)}<M<X_{(j)})=\sum_{k=i}^{j-1}\binom{n}{k}\left(\frac{1}{2}\right)^{k}\left(\frac{1}{2}\right)^{n-k}$$

我们知道 $M=m_{0.5}$ 是 0.5 分位数. 类似地,可得到一般 p 分位数 m_p 的 $100(1-\alpha)\%$ 的置信区间 $(X_{(i)},X_{(j)})$,这里 i,j 满足

$$1-\alpha=P(X_{(i)}<m_p<X_{(j)})=\sum_{k=i}^{j-1}\binom{n}{k}p^{k}(1-p)^{n-k}$$

符号检验可用来检验成对数据的比较. 一个简单例子是比较不同材料制成的左右两双鞋的磨损程度. 假定 $(X_1,Y_1),\cdots,(X_n,Y_n)$ 为 n 对数据. 看是否 X_i 和 Y_i 大体相同. 这就归结到对样本 $Z_1,\cdots,Z_n(Z_i=X_i-Y_i,\quad i=1,\cdots,n)$ 的符号检验问题,即检验其中位数是否为零.

符号检验不需任何对总体分布的假设,简单易懂,缺点是没有利用数据大小的全部信息. 以后要介绍的其它非参数统计方法,则注意到了这一点.

2.3 Cox-Stuart 趋势检验

在各种统计结果中,特别是涉及经济、人口、环境、卫生等随着时间变化的统计数据,人们往往关心变化的趋势,比如,收入是否下降了,环境是否变坏了,气候是否变暖了等问题. 给定一组数据后,如何看其趋势呢? 最常见的参数方法是用线性回归拟合一条直线,再看其是否上升. 然而,单调的趋势不一定是线性的,也不一定能由一个显函数来表达. 这里我们来考虑一个简单的非参数方法. 直观上,能通过前后数据的比较来看一组数据是否有单调的趋势. 我们可以选许多对数据,每一对由前后两个不同时间的数据组成,它们的间隔应尽可能地远,因为

个总体上升或下降的数据在局部上可能有小的不规则的波动，这些数据对应等距，以便它们的差为同分布的. 因此每对中两数的距离也不能太大而使得对的数目太少. 为此 Cox 和 Stuart 于 1955 年提出了基于符号检验的非参数方法. 假设有 n 个数据 $X_1,\cdots,X_n$. 我们想看是否随着下标有上升或下降的趋势. 换句话说，就是下列三个检验问题之一：

1. H_0：无趋势 $\leftrightarrow H_1$：有升或降的趋势；
2. H_0：无上升趋势 $\leftrightarrow H_1$：有上升趋势；
3. H_0：无下降趋势 $\leftrightarrow H_1$：有下降趋势.

我们的数据对可取为 $(X_1,X_{1+c}),\cdots,(X_{n-c},X_n)$，这里当 n 为偶数时，$c=\frac{n}{2}$，当 n 为奇数时，$c=\frac{n+1}{2}$. 易见，当 n 为偶数时共有 c 对；而当 n 为奇数时共有 $c-1$ 对，这时数据 X_c 没有配上对. 令 $D_i=X_i-X_{i+c}$；S^+ 为 $\{D_i\}$ 中正号的数目，而 S^- 为负号的数目. 当没有趋势时，S^+ 或 S^- 为 $p=0.5$ 的二项分布. 显然，如果 S^+ 大（或 S^- 小），则可能有下降趋势，而如果 S^- 大（或 S^- 小），则可能有上升趋势. 相应于上面三个检验问题，分别取检验统计量

$$1.\ K=\min(S^+,S^-);\quad 2.\ K=S^-;\quad 3.\ K=S^-$$

检验过程和前面的符号检验完全一样. 当 K 太小时，我们拒绝原假设.

下面以一例来说明这个 Cox-Stuart 检验.

美国国家宇航局（NASA）自 1966 至 1984 年的科研和发展经费按时间顺序为（单位为千万美元）：

5.9	5.4	4.7	4.3	3.8	3.4	3.4	3.3	3.3	3.3
3.7	3.9	4.0	4.2	4.9	5.2	6.0	6.7	7.0	

我们有 $n=19,c=10,S^-=4,S^-=5$. 如考虑上面检验 1，即 H_0：无趋势 $\leftrightarrow H_1$：有趋势，取检验统计量 $K=\min(S^+,S^-)$. K 的值为 $k=4$；p 值为 $2P(K\leqslant k)=1$. 因此，即使取水平 $\alpha=1$ 也不拒绝原假设. 但如果我们只取自 1970 至 1984 的数据，并考虑上面检验 2，即 H_0：无上升趋势 $\leftrightarrow H_1$：有上升趋势. 有 $n=15,c=8,S^+=0,S^-=7$. p 值为 0.0078. 于是，对所有水平 $\alpha>0.0078$ 都可拒绝原假设. 这和前面的结果似乎矛盾. 实际上，如果我们仔细观查原数据或散点图的话，就可以看出数据是先下降后上升的. 因此用 Cox-Stuart 方法就检验不出有任何趋势了. 此例也说明预先对数据进行初步分析（如散点图）的好处.

2.4 随机游程检验

通常所说的随机性是指样本中所有数据都可看成是独立同分布的观察值，上面提到的有升降趋势的数据不是随机的，有周期性变化的数据也不是随机的. 当数据正相关时，大的或小的数据往往有聚在一起的倾向. 负相关时，则正相反. 这一节，我们主要考虑二元数据的观察值（比如 Bernoulli 试验的结果），它们总可以用 0 和 1 来表示. 在一个随机的观察值序列中，0 或 1 的集中度有一定的范围，我们因此引进游程的概念来描述这种集中程度. 在一个由 0 和 1 组成的序列中，一串不间断的 0 或 1 称为一个游程（run），一个游程中数字“0”或“1”的个数，称为

该游程的长度.游程个数 R 太多，则说明 0 和 1 不集中或游程太短(负相关)；如游程个数太少，则说明 0 和 1 较集中或游程太长(正相关).通过上面的分析，我们知道随机性假设的拒绝域应为 $\{R \leqslant c_1\} \cup \{R \geqslant c_2\}$，$(c_1 < c_2)$.比如 0 和 1 的序列

$$1\quad 1\quad 0\quad 0\quad 0\quad 1\quad 1\quad 0\quad 0\quad 1\quad 1\quad 1\quad 0$$

有 6 个游程，用 $R=6$ 表示；其中有 3 个是由 0 组成的(长度分别为 3,2,1)，用 $m=6$ 表示数字 0 的个数；3 个是由 1 组成的(长度分别为 2,2,3)，用 $n=7$ 表示数字 1 的个数.在零假设下(随机性)，R 的分布依赖于出现 1 的未知概率 p.但是，在给定 m 和 n 的条件下，R 的任何一种可能的概率都是 $\dfrac{1}{\binom{N}{n}}$，$(N=m+n)$.因而有

$$P(R=2k)=\frac{2\binom{m-1}{k-1}\binom{n-1}{k-1}}{\binom{N}{n}}$$

及

$$P(R=2k+1)=\frac{\binom{m-1}{k-1}\binom{n-1}{k}+\binom{m-1}{k}\binom{n-1}{k-1}}{\binom{N}{n}}$$

这个表示很简单实用，并且 Swed 和 Eisenhart 于 1943 年依此构造了 R 的零分布表(见附表 4).对于大样本来说，当 $n\to\infty$ 而 $\frac{m}{n}\to\gamma$ 时，则有

$$\frac{R-\dfrac{2m}{1-\gamma}}{\sqrt{\dfrac{4\gamma m}{(1+\gamma)^3}}}$$

渐近趋于标准正态分布(证明见[1])，于是当样本容量很大时，可近似地取临界值为

$$c_1=\frac{2mn}{m+n}\left(1+\frac{Z_{\frac{\alpha}{2}}}{\sqrt{m+n}}\right),\quad c_2=1+\frac{2mn}{m-n}\left(1-\frac{Z_{\frac{\alpha}{2}}}{\sqrt{m+n}}\right)$$

上面我们仅讲了二元数据随机性的游程检验，而实际中遇到的数据未必都是二元数据，此时，我们就要把数据转化成二元数据，以利用上面的游程检验.事实上，如取 $Y_i=I(X_i-X_{med}>0)$，则可以把检验 $X_1,\cdots,X_n$ 的随机性问题转化成检验 $Y_1,\cdots,Y_n$ 的随机性问题.当然，这种转化不是完全等价的，这是 Mood 于 1940 年给出的，有兴趣的读者可见 Ann. Math. Statist.,1940(11):367—392.我们称此种方法为中位数法.

例 2.1 对某型号电缆进行耐压试验，测得其 20 根的数据如下：

156.0,255.5,132.0,246.7,867.9,86.4,610.4,125.7,150.4,117.6
201.9,207.2,189.8,585.8,153.1,565.4,511.0,567.0,222.3,141.5

根据这些数据能否认为这些电缆受到了非随机因素的干扰?或者说,能否认为生产这种电缆的机器不正常?

对于本例,我们利用上述中位数法,计算样本中位数为 204.6,相应的 Y 样本为:

$$0,1,0,1,1,0,1,0,0,0,0,1,0,1,0,1,1,1,1,0$$

则 $m=n=10, R=13$,对于 $\alpha=0.1$,由附表 4 查得 $c_1=6, c_2=16$,因为 $6<13<16$,则认为这些数据符合随机性假设.

2.5 阅读知识

前面我们讲过样本中位数是总体中位数的渐近无偏估计. 实际上还可以证明,任给 $0<p<1$,设 $X_1,\cdots,X_n$ 为来自 $F(x)$ 的独立同分布样本,$F(x)$ 在其 p 分位数 ξ_p 处连续且存在密度函数 $f(x)$,并有 $f(\xi_p)>0$. 则样本 p 分位数 $m_p \xrightarrow{\text{a.s.}} \xi_p, n\to\infty$.

事实上,因为 $f(\xi_p)>0$,则 ξ_p 唯一,且对于任给的 $\varepsilon>0$ 有

$$F(\xi_p-\varepsilon)<p<F(\xi_p+\varepsilon)$$

由强大数定律知

$$\lim_{n\to\infty}\frac{\#\{i: X_i<\xi_p-\varepsilon, i=1,\cdots,n\}}{n}\overset{\text{a.s.}}{=}F(\xi_p-\varepsilon)<p$$

(这里的符号"#"表示计数的意思),所以

$$P(\text{对于充分大的 } n, X_1,\cdots,X_n \text{ 中小于 } \xi_p-\varepsilon \text{ 的个数不超过 } np-1)=1$$

又由于 $[np]>np-1$ 及 m_p 的定义,则知

$$P(\text{对于充分大的 } n, m_p\geqslant\xi_p-\varepsilon)=1$$

同理可证

$$P(\text{对于充分大的 } n, m_p\leqslant\xi_p+\varepsilon)=1$$

即 $m_p\xrightarrow{\text{a.s.}}\xi_p, n\to\infty$.

下面我们看一看有关 ξ_p 的置信区间与置信限,下设 ξ_p 是唯一的.

首先求形如 $X_{(r)}$ 的置信上限,即对于给定的 $\alpha>0$,求 r,使

$$P(X_{(r)}\geqslant\xi_p)=1-\alpha$$

由 §1.4 知 $X_{(r)}$ 的确切分布,故 ξ_p 的 $1-\alpha$ 的置信上限 $X_{(r)}$ 应满足

$$\sum_{k=r}^{n}\binom{n}{k}p^k(1-p)^{n-k}=\alpha$$

或者

$$r\binom{n}{r}\int_0^p t^{r-1}(1-t)^{n-r}\mathrm{d}t=\alpha$$

当 n 不很大时,可以通过二项分布表(见附表 1)查得 r 的值. 但是应注意到,对于事先给定的 $\alpha>0$,不一定恰好有一个正整数 r 满足上式,而可能存在一个正整数 r_0,使得

$$\sum_{k=r_0+1}^{n}\binom{n}{k}p^k(1-p)^{n-k} < \alpha < \sum_{k=r_0}^{n}\binom{n}{k}p^k(1-p)^{n-k}$$

此时，有两种处理方法：一是看上式两端哪一个与α接近，就取那个正整数；二是用下面的随机化方法，令

$$\delta = \frac{\sum_{k=r_0}^{n}\binom{n}{k}p^k(1-p)^{n-k}-\alpha}{\binom{n}{r_0}p^{r_0}(1-p)^{n-r_0}}$$

显然$0<\delta<1$. 如任取一个$U(0,1)$上的随机数t，则r的取法如下：

$$r=\begin{cases} r_0+1, & t\leqslant\delta \\ r_0, & t>\delta \end{cases}$$

验证一下知，这样得到的置信上限$X_{(r)}$其置信度的确为$1-\alpha$.

当n很大时，由于二项分布收敛于正态分布，故总体分位数ξ_p的置信上限$X_{(r)}$可以近似地写成

$$r \doteq np+0.5+Z_\alpha\sqrt{np(1-p)}$$

关于ξ_p的形如$X_{(s)}$的置信下限，可类似求得，望读者自己补上.

有了ξ_p的置信上限及置信下限之后，则可以近似地得到ξ_p的置信区间. 首先分别求出ξ_p的$1-\frac{\alpha}{2}$的置信上下限$X_{(s)}, X_{(r)}$，则其$1-\alpha$的置信区间可近似地取为$[X_{(r)}, X_{(s)}]$. 事实上，这样得到的置信度不低于$1-\alpha$.

$$\begin{aligned} & P(X_{(r)}\leqslant\xi_p\leqslant X_{(s)}) \\ \geqslant & P(X_{(r)}\leqslant\xi_p)+P(\xi_p\leqslant X_{(s)})-1 \\ = & 1-\alpha \end{aligned}$$

由此可见，这样得到的置信区间是比较保守的. 当然，精确的求法也是存在的，感兴趣的读者可参见[1].

2.6 习　　题

1. 设$X_1,\cdots,X_n$为n个来自某连续分布函数$F(x)$的 iid 样本，试求满足

(1) $P(X_{(1)}\leqslant M\leqslant X_{(n)})\geqslant 0.95$;

(2) $P(F(X_{(n)})-F(X_{(1)})\geqslant 0.5)\geqslant 0.95$

的最小的n(M为中位数).

2. 设分布函数$F(x)$连续，M为其中位数，$X_1,\cdots,X_n$为来自$F(x)$的 iid 样本，如$(X_{(r)}, X_{(n-r)})$ $(r<\frac{n}{2})$是M的$1-\alpha$置信区间，则α满足

$$\alpha=(0.5)^{n-1}\sum_{k=0}^{r-1}\binom{n}{k}=2n\binom{n-1}{r-1}\int_0^{\frac{1}{2}}x^{n-r}(1-x)^{r-1}\mathrm{d}x$$

3. 设$X_1,\cdots,X_n\overset{\text{iid}}{\sim}F(x)$，$\xi_p$为$F(x)$的$p$分位数，记

$p_<=P(X_1<\xi_p)$,　$p_>=P(X_1>\xi_p)$,　$S_<=\#\{i:X_i<\xi_p\}$,　$S_>=\#\{i:X_i>\xi_p\}$

证明

(1) $E(S_{<}-S_{>})=n(p_{<}-p_{>})$, $\mathrm{Var}(S_{<}-S_{>})=n(p_{<}+p_{>}-(p_{<}-p_{>})^2)$;

(2) 如 $p_{<}+p_{>}=1$,则以概率1保证 $S_{<}-S_{>}$ 的奇偶性与 n 相同.

4. 试证明,由随机化方法得到的中位数的置信上限的置信度的确为 $1-\alpha$.

5. 现从某个单位中随机地抽取9名职工,其日收入为(单位:元):

38.3,41.2,47.0,36.5,39.1,38.9,38.3,40.2,37.9

(1) 试求其中位数的估计;

(2) 试求中位数的95%的置信区间.

6. 设有甲、乙两地,甲在乙地的东边,现在甲地饲养10只信鸽,过一段时间后,送到乙地放飞,并测其消失时的飞行方向分别为向东偏北20°,35°,350°,120°,85°,345°,80°,320°,280°,85°. 显然,由0°—90°和270°—360°中的数据表明,该信鸽的飞行方向偏东,否则,就偏西.

(1) 求其中位数;

(2) 试求中位数的95%置信区间;

(3) 试用符号统计量检验这批信鸽飞行方向是否偏东?并求出其 p 值.

7. 现有新、旧两个小麦品种,把他们分别同时种在8块地上做试验,测得其产量为(单位:公斤/亩):

土地	1	2	3	4	5	6	7	8
新品种	209	200	177	169	159	187	169	198
旧品种	151	168	147	164	166	176	169	188

(1) 试求新旧品种小麦亩产量中位数的估计;

(2) 试用符号统计量检验新品种小麦是否优于旧品种?

8. 在某一地区,人们测量到某种类的成年猴的平均体重为8.41公斤,而在另一地区人们观测到此种成年猴的体重为:

8.30 9.50 9.60 8.75 8.40 9.10 9.25 9.80
10.05 8.15 10.00 9.60 9.80 9.20 9.30

从这组数据,我们能否说这组猴的体重的中位数大于8.41公斤?并求出其 p 值.

9. 在某一地区,现从有吸毒史的病人中抽取15人,询问其第一次吸毒时的年龄(岁)如下:

22,24,37,28,15,14,22,16,18,17,23,16,20,18,15

(1) 由此能否说本地区吸毒人第一次吸毒的年龄中位数为20?并求其 p 值;

(2) 并求出中位数的95%的置信区间.

10. 在某一学校,随机地抽取20名"下海"的大学生,询问其"下海"的原因是不是为了挣钱以减轻家里的负担,其中有6人回答"是",请问这些数据能否说明"下海"学生挣钱的目的在于减轻家里的负担,其 p 值为多少?

11. 我国1949—1983年间大豆总产量如下(单位:万吨)

年份	1949	1950	1951	1952	1953	1954	1955	1956	1957
产量	509	744	863	952	993	908	912	1024	1005
年份	1958	1959	1960	1961	1962	1963	1964	1965	1966
产量	867	876	639	621	651	691	787	614	827
年份	1967	1968	1969	1970	1971	1972	1973	1974	1975
产量	827	804	763	871	861	645	837	747	724
年份	1976	1977	1978	1979	1980	1981	1982	1983	
产量	664	726	757	746	794	933	903	976	

试问大豆的产量是否有上升的趋势？并求其 p 值.

12. 于1973年美国联邦保险公司提交给国会的年度报告中指出，棉花入保的数量如下：

年份	1948	1949	1950	1951	1952	1953
产量	19479	26667	33969	57715	38086	38434
年份	1954	1955	1956	1957	1958	1959
产量	24196	19319	29975	25451	20410	19910
年份	1960	1961	1962	1963	1964	1965
产量	15628	15375	2132	26526	24865	21152
年份	1966	1967	1968	1969	1970	1971
产量	23458	25774	32646	31786	24821	19593

试回答：棉花入保量是否在逐年下降？

13. 试证明游程检验统计量的零分布的期望与方差如下：

$$E(R) = 1 + \frac{2mn}{m+n}$$

$$\mathrm{Var}(R) = \frac{2mn(2mn - m - n)}{(m+n)^2(m+n-1)}$$

14. 在某一电话亭，观测打电话人的性别（以 M 表示男性，F 表示女性）依次如下：$FFFMFFMMMFFFFFFM$，试问男女的出现是否为随机的？

15. 一洗发剂生产厂家的质检科要求每瓶洗发剂的平均重量为12液量盎司. 现从一台机器中随机抽取20瓶，测其重量如下：

12.9　12.5　12.2　12.3　11.5　11.8　11.7　12.2　12.4　12.6
12.5　12.8　11.8　11.5　11.6　12.7　12.6　12.7　12.8　12.2

试验证这台机器多灌少灌是不是随机的.

第三章 对称分布的单样本问题

3.1 引 言

前面一章讲的非参数方法对总体未作任何要求，因而适应性很广．但是，如果我们已知总体分布的一些性质而不利用，就会浪费许多有用的信息．最常见的就是分布的对称性．有了对称性，我们可用更有效的方法以利用数据中的尽可能多的信息．很多数据看起来并不对称，但在一些变换下(如指数变换)可成为对称的，或大体上是对称的．因此对称性并不是一个很强的条件．基于对称性假设的方法有广泛的应用性．在非参数方法中，我们感兴趣的位置参数主要是中位数；而在参数方法中则为均值．如果分布对称而且中位数唯一，这二者就是等同的(可称之为中心)，因而可以比较参数方法和非参数方法在不同条件下的优劣．

在本章中，我们主要考虑连续的对称分布．称一个连续分布函数 $F(x)$ 关于原点对称，如果 $\forall\, x \in R, F(-x) = 1 - F(x)$．用概率表示为，设 $X \sim F(x)$，则 $\forall\, x \in R$，

$$P(X < -x) = P(X > x)$$

设 θ 为一实数，称随机变量 X 或分布函数 $F(x)$ 关于 θ 对称，如果随机变量 $X - \theta$ 或者分布函数 $F(x + \theta)$ 关于原点对称，且 θ 称为对称中心．

用 Ω_0 表示所有连续的中位数等于零($F(0) = \frac{1}{2}$) 的分布所组成的族，用 Ω_S 表示 Ω_0 中的对称分布类．即

$$\Omega_S = \{F : F \in \Omega_0; F(-x) = 1 - F(x)\}$$

关于对称分布，由简单的概率运算，可得以下性质(有些上面已提到)．

定理 3.1 随机变量 X 关于 θ 对称当且仅当 $X - \theta$ 和 $\theta - X$ 依分布相等．

证明 必要性．

因为 X 关于 θ 对称，则 $X - \theta$ 关于原点对称，即

$$\forall\, x \qquad P(X - \theta < x) = P(X - \theta > -x) = P(\theta - X < x)$$

即 $X - \theta$ 与 $\theta - X$ 依分布相等．

充分性．

任给 x，因为 $X - \theta$ 与 $\theta - X$ 依分布相等，所以

$$P(X - \theta < x) = P(\theta - X < x) = P(X - \theta > -x)$$

即 X 关于 θ 对称． □

推论 3.1 如果随机变量 X 关于 θ 对称，且其期望存在，则期望等于 θ．

推论 3.2 对称分布的对称中心必唯一．

上面两个推论的证明留作习题．

定理 3.2　如果随机变量 X 关于 θ 对称，则对称中心 θ 是总体中位数之一.

证明　因为 X 关于 θ 对称，所以

$$\forall\, x,\qquad P(X-\theta<x)=P(X-\theta>-x)$$

特别地，取 $x=0$，则

$$P(X<\theta)=P(X>\theta)P(X<\theta)\leqslant\frac{1}{2}$$

下证 $P(X\leqslant\theta)\geqslant\frac{1}{2}$. 反证，如 $P(X\leqslant\theta)<\frac{1}{2}$，则

$$P(X>\theta)=P(X<\theta)=1-P(X\leqslant\theta)>\frac{1}{2}$$

这与上面给论矛盾，综合两者，有

$$P(X<\theta)\leqslant\frac{1}{2}\leqslant P(X\leqslant\theta)$$

即 θ 是 X 的一个中位数.　　□

3.2　秩及有关分布

在样本 $X_1,\cdots,X_n$ 中，X_i 如果是第 R_i 个最小的，即 $X_i=X_{(R_i)}$（第 R_i 个顺序统计量），则我们称 R_i 为 X_i 的秩. 显然，$R_i=\sum_{j=1}^{n}I(X_j\leqslant X_i)$. 令 $R=(R_1,\cdots,R_n)$. 于是 R 为样本的一个统计量. 凡是由秩产生的统计量都称为秩统计量. 在样本是随机(iid)的时候，$R=(R_1,\cdots,R_n)$ 取 $(1,\cdots,n)$ 的任意 $n!$ 个排列之一的概率都是一样的，即 $\frac{1}{n!}$. 换句话说，R 在由 $(1,\cdots,n)$ 的所有排列组成的空间上是均匀分布. 于是有了下面的定理：

定理 3.3　对于 iid 样本，对 $(1,\cdots,n)$ 的任一排列 $(i_1,\cdots,i_n)$ 有

$$P(R=(i_1,\cdots,i_n))=\frac{1}{n!}$$

证明　记

$$\mathscr{R}=\{(j_1,\cdots,j_n)\text{ 是 }(1,\cdots,n)\text{ 的一个排列}\}$$

则 $\mathscr{R}$ 中共有 $n!$ 个元素.

$$\begin{aligned}&P(R=(i_1,\cdots,i_n))\\=&P((X_1,\cdots,X_n)=(X_{(i_1)},\cdots,X_{(i_n)}))\\=&P(X_{d_1}<X_{d_2}<\cdots<X_{d_n})\end{aligned}$$

其中 d_i 表示数 i 在 $(i_1,\cdots,i_n)$ 中由小到大排列的位次，即，$X_{(i)}=X_{d_i}$. 因为 $X_1,\cdots,X_n$ 是 iid 的，所以 $(X_1,\cdots,X_n)$ 与 $(X_{d_1},\cdots,X_{d_n})$ 同分布，即

$$\begin{aligned}&P(R=(i_1,\cdots,i_n))\\=&P(X_{d_1}<X_{d_2}<\cdots<X_{d_n})\\=&P(X_1<X_2<\cdots<X_n)\\=&P(R=(1,\cdots,n))\end{aligned}$$

则 R 的分布与其取值无关，又因为 $\mathscr{R}$ 中共有 $n!$ 个元素，所以

$$P(R = (i_1, \cdots, i_n)) = \frac{1}{n!} \qquad \square$$

上面定理说的是 $R_1, \cdots, R_n$ 的联合分布. 类似地，每一个 R_i 在空间 $1, \cdots, n$ 上有均匀分布（在每一点的概率为 $\frac{1}{n}$）；每一对 (R_i, R_j) 在空间 $\{(r,s): r,s = 1, \cdots, n, r \neq s\}$ 上有均匀分布（在每一点的概率为 $\frac{1}{n(n-1)}$）. 以推论的形式有：

推论 3.3　对于 iid 样本，对任意 $r, s = 1, \cdots, n, r \neq s$ 及 $i \neq j$，

$$P(R_i = r) = \frac{1}{n}$$

$$P(R_i = r, R_j = s) = \frac{1}{n(n-1)}$$

读者还很容易得到下面的另一推论.

推论 3.4　对于 iid 样本，对任意 $r, s = 1, \cdots, n, r \neq s$ 及 $i \neq j$，

$$E(R_i) = \frac{n+1}{2}$$

$$\mathrm{Var}(R_i) = \frac{(n+1)(n-1)}{12}$$

$$\mathrm{Cor}(R_i, R_j) = -\frac{n+1}{12}$$

理论上，用类似方法可得到 $(R_1, \cdots, R_k)$，$1 \leqslant k \leqslant n$ 的所有可能的联合分布. 有兴趣的读者可试一试. 从上面定理可见，对于 iid 样本，秩统计量的分布和原来的总体分布没有关系（distribution-free）. 我们也未对总体分布作任何假设.

前面的秩统计量只考虑了样本点的大小而未考虑其绝对值的大小，但其绝对值的大小有时是很重要的. 例如对数据 $-0.21, -0.2, -0.13, -0.01, 0, 15, 50, 100, 150$ 来说，0 是中位数，有正号和有负号的数目一样；如果只看秩，而不看原来数据，给人的印象是一个很对称的样本. 但实际上则不然. 问题出在数值的绝对值的大小没有考虑进去. 现在引进 Wilcoxon 符号秩统计量，用 W^+ 来表示. 我们把样本的绝对值 $|X_1|, \cdots, |X_n|$ 排序. 其顺序统计量为 $|X|_{(1)}, \cdots, |X|_{(n)}$. 用 R_j^+ 表示 $|X_j|$ 在绝对值样本中的秩，即 $|X_j| = |X|_{(R_j^+)}$. 我们还用 $S(x)$ 表示符号函数 $I(x > 0)$，它在 $x > 0$ 时为 1，否则为 0. 为方便起见，我们引入反秩（antirank）的概念. 反秩 D_j 是由 $|X_{D_j}| = |X|_{(j)}$ 定义的. 我们还用 W_j 表示与 $|X|_{(j)}$ 相应的原样本点的符号函数，即 $W_j = S(X_{D_j})$，且称 $R_j^+ S(X_j)$ 为符号秩统计量. Wilcoxon 符号秩统计量定义为

$$W^+ = \sum_{j=1}^{n} jW_j = \sum_{j=1}^{n} R_j^+ S(X_j)$$

它是正的样本点按绝对值所得的秩的和. 为说明这些概念，有如下例子.

例 3.1　如样本值为：7，9，−2，8，−10，1，则相应的统计量值为

X_1	X_2	X_3	X_4	X_5	X_6
7	9	−2	8	−10	1
$\|X\|_{(3)}$	$\|X\|_{(5)}$	$\|X\|_{(2)}$	$\|X\|_{(4)}$	$\|X\|_{(6)}$	$\|X\|_{(1)}$
$R_1^+=3$	$R_2^+=5$	$R_3^+=2$	$R_4^+=4$	$R_5^+=6$	$R_6^+=1$
$W_3=1$	$W_5=1$	$W_2=0$	$W_4=1$	$W_6=0$	$W_1=1$
$D_3=1$	$D_5=2$	$D_2=3$	$D_4=4$	$D_6=5$	$D_1=6$

显然 $W^+=3+5+4+1=13$.

下面我们介绍上述有关统计量的一些性质，为下一节的 Wilcoxon 符号秩检验作准备. 假设 $F(x-\theta)\in\Omega_S$，通常的零假设为 $H_0:\theta=0$. 按照本章的记号，我们有下面 3 个定理.

定理 3.4　如果零假设 $H_0:\theta=0$ 成立，则 $S(X_1),\cdots,S(X_n)$ 独立于 $(R_1^+,\cdots,R_n^+)$.

证明　事实上，因为 $(R_1^+,\cdots,R_n^+)$ 是 $|X_1|,\cdots,|X_n|$ 的函数，而出自随机样本的 $(S(X_i),|X_i|)$，$i=1,\cdots,n$ 是互相独立的对子，因此我们只要证明 $S(X_i)$ 和 $|X_i|$ 是互相独立的即可. 事实上，

$$P(S(X_i)=1,|X_i|\leqslant x)=P(0<X_i\leqslant x)=F(x)-F(0)=F(x)-\frac{1}{2}$$
$$=\frac{2F(x)-1}{2}=P(S(X_i)=1)P(|X_i|\leqslant x)\qquad\square$$

下面的定理 3.5 和定理 3.4 平行，读者可自己验证.

定理 3.5　如果零假设 $H_0:\theta=0$ 成立，则 $S(X_1),\cdots,S(X_n)$ 独立于 $(D_1,\cdots,D_n)$.

定理 3.6　如果零假设 $H_0:\theta=0$ 成立，则 $W_1,\cdots,W_n$ 是独立同分布的，其分布为 $P(W_i=0)=P(W_i=1)=\frac{1}{2}$.

证明　令 $D=(D_1,\cdots,D_n)$，$d=(d_1,\cdots,d_n)$，

$$\begin{aligned}&P(W_1=w_1,\cdots,W_n=w_n)\\&=\sum_d P(S(X_{D_1})=w_1,\cdots,S(X_{D_n})=w_n|D=d)P(D=d)\\&=\sum_d P(S(X_{d_1})=w_1,\cdots,S(X_{d_n})=w_n)P(D=d)\\&=\left(\frac{1}{2}\right)^n\sum_d P(D=d)=\left(\frac{1}{2}\right)^n\end{aligned}$$

因此有 $P(W_1,\cdots,W_n)=\prod_{i=1}^{n}P(W_i=w_i)$ 及 $P(W_i=w_i)=\frac{1}{2}$.　$\square$

3.3　Wilcoxon 符号秩检验

在这一节，我们考虑前面提到的检验问题. 假定 iid 样本来自对称分布总体，即 $X_1,\cdots,X_n\sim F(x-\theta)$，$F(x)\in\Omega_S$. 需要检验的是 $H_0:\theta=0$. 这里用前面定义的 Wilcoxon 符号秩统计量

W^+ 来检验. 直观上,W^+ 如太大或太小都对零假设提出挑战. 要进行更精确的检验,需要 W^+ 的分布,以对其实际值 t 求 p 值($P(|W^+|>t)$ 或 $P(W^+>t)$ 或 $P(W^+<t)$ 依不同的 H_1 而定). 因为 $W^+=\sum_{j=1}^{n} jW_j$,由前面的关于 W_i 分布的定理,W_j 是独立同分布的(二项分布 $B(1,\frac{1}{2})$)随机变量的线性组合,显然独立于总体分布. 虽然我们无法用一个简单表达式写出 W^+ 的分布,但 W^+ 的分布实际上是很简单的.

下面一个例子表明如何对简单情况直接算出 W^+ 的值及有关的概率.

例 3.2 假定 $n=3$,则所有可能的绝对值的秩为 1,2,3,所有可能的正负号的组合为 $2^3=8$ 种. 在零假设下,每种可能的概率为 $p=\frac{1}{8}$. 用下表来表示所有可能的组合及相应 W^+ 的值. 左边的 1,2,3 为所有可能的秩.

1	−	+	−	−	+	+	−	+
2	−	−	+	−	+	−	+	+
3	−	−	−	+	−	+	+	+
W^+	0	1	2	3	3	4	5	6

上表中,实现每一列的概率为$\frac{1}{8}$. 可见,$P(W^+=k)=\frac{1}{8}$,$k=0,1,2,4,5,6$;及 $P(W^+=3)=\frac{2}{8}=\frac{1}{4}$. 如果 H_1 为 $W^+>t_0$,而 W^+ 的观察值为 5,则 p 值为 $P(W^+\geqslant 5)=P(W^+=5)+P(W^+=6)=\frac{1}{4}$. 这个 p 值太大,不能拒绝零假设. 事实上,因为这个样本太小,对任何 W^+ 的值都不大可能拒绝零假设. 换句话说,对于如此小的样本,作任何推断的证据都不足.

现在我们给出在一般情况下计算 W^+ 的零分布的方法,该方法是编写简单计算机程序的基础. 首先找出 W^+ 的矩母函数 $M(t)$. 注意,下面的期望是对零假设而言. 对任意的 j,有

$$E(\exp(tjW_j))=\frac{1}{2}\exp(0)+\frac{1}{2}\exp(tj)=\frac{1}{2}(1+\exp(tj))$$

因而

$$M(t)=E(\exp(tW^+))=E(\exp(t\sum_j jW_j))$$

$$=\prod_j E(\exp(tjW_j))=\frac{1}{2^n}\prod_{j=1}^{n}(1+e^{tj})$$

按矩母函数的性质,如果 $M(t)=a_0+a_1e^t+a_2e^{2t}+\cdots$,则对 $j=0,1,\cdots$,有 $P(W^+=j)=a_j$.

当 $n=2$ 时,

$$M(t)=\frac{(1+e^t)(1+e^{2t})}{2^2}=\frac{1+e^t+e^{2t}+e^{3t}}{4}$$

我们有下表:其第一行是 $M(t)$ 的指数幂,第二行是相应于第一行的系数(差一个因子$\frac{1}{4}$).

0	1	2	3
1	1	1	1

由此可得

$$P(W^+=1)=P(W^+=2)=P(W^+=3)=\frac{1}{4}$$

当 $n=3$ 时,

$$M(t)=\frac{(1+e^t)(1+e^{2t})(1+e^{3t})}{2^3}$$

我们有下表:其第一行是$M(t)$的指数幂,第二行是前表的结果,是由第三个因子$(1+e^{3t})$中第一项(即1)乘头两项而得.第三行又是第二行的重复,但右移三位,是由第三个因子中第二项(即e^{3t})乘前面项而得,因是三次幂,故位移三位.第四行是第二三两行的和,即第一行的相应的系数(差一个因子$\frac{1}{8}$).

0	1	2	3	4	5	6
1	1	1	1			
			1	1	1	1
1	1	1	2	1	1	1

由此得到和上例一样的结果:即$P(W^+=k)=\frac{1}{8}$, $k=0,1,2,4,5,6$及$P(W^+=3)=\frac{2}{8}=\frac{1}{4}$.类似地,可有相对于$n=4$时的表:

0	1	2	3	4	5	6	7	8	9	10
1	1	1	2	1	1	1				
				1	1	1	2	1	1	1
1	1	1	2	2	2	2	2	1	1	1

它只用了前一个表的结果.我们因而可以很容易地编出个短小的子程序,以对任意的n值去计算W^+的概率.当然,许多书上提供了W^+的分布表.但在计算机上还是用子程序计算较方便,本书的附表3给出了部分零分布表.

除了用计算机算W^+的分布之外,还可用正态近似,此时我们需要其均值和方差.由上节的定理,易得(这里和以下的期望和方差都是在零假设下取的)

$$E(W^+)=E\sum_j(jW_j)=\frac{1}{2}\sum_j j=\frac{n(n+1)}{4}$$

$$\mathrm{Var}(W^+)=\mathrm{Var}\sum_j(jW_j)=\frac{1}{4}\sum_j j^2=\frac{n(n+1)(2n+1)}{24}$$

由中心极限定理,在n大时,可近似地认为

$$\frac{W^{-} - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \xrightarrow{\mathscr{L}} N(0,1)$$

而对于单边检验 $H_0:\theta = 0 \leftrightarrow H_1:\theta > 0$ 及水平 α，检验的临界值为

$$c \approx \frac{n(n+1)}{4} + 0.5 + Z_\alpha \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

除正态近似之外，还有一些其它的对于概率 $P(W^{+} \leqslant k)$ 的近似，可参阅有关资料.

对于上面的正态近似，也可以如下表示：当 n 很大时，有

$$P(W^{+} \leqslant k) \approx \Phi\left(\frac{k - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}\right)$$

如利用连续性修改(continuity correction)，则有如下的近似

$$P(W^{+} \leqslant k) \approx \Phi\left(\frac{k + 0.5 - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}\right)$$

其中 $\Phi(\cdot)$ 为标准正态分布函数. 我们对上面的近似进行了数字比较，见下表：

当 $n = 10$ 时与 $n = 20$ 时分别为

k	3	5	11	14	17
精确值	0.0049	0.0098	0.0527	0.0967	0.1611
连续修改	0.0072	0.0124	0.0515	0.0926	0.1541
无连续修改	0.0063	0.0109	0.0463	0.0844	0.1423

和

k	37	43	61	70	76
精确值	0.0047	0.0096	0.0527	0.1012	0.1471
连续修改	0.0059	0.0108	0.0522	0.0989	0.1437
无连续修改	0.0055	0.0103	0.0502	0.0957	0.1394

从上两个表可以看出，近似效果还是不错的，尤其是连续性修改之后.

例 3.3 在 8 块土地上同时试种新、旧两种小麦，而我们知道旧小麦的平均亩产量为 180(公斤 / 亩)，而新小麦产量为：209，200，179，230，170，195，210，155. 试问：新产品小麦有无推广的必要？

对于此题，我们不妨假设小麦产量的分布为对称的，则问题归结为如下的检验：

$$H_0:\theta = 180 \leftrightarrow H_1:\theta > 180$$

此时我们可以用 Wilcoxon 符号秩检验. 令 $Y_i = X_i - 180$，$i = 1,\cdots,8$，经计算知

Y_1	Y_2	Y_3	Y_4
29	20	-1	50
$S_1=1$	$S_2=1$	$S_3=0$	$S_4=1$
$R_1^+=6$	$R_2^+=4$	$R_3^+=1$	$R_4^+=8$
Y_5	Y_6	Y_7	Y_8
-10	15	30	-25
$S_5=0$	$S_6=1$	$S_7=1$	$S_8=0$
$R_5^+=2$	$R_6^+=3$	$R_7^+=7$	$R_8^+=5$

则 Wilcoxon 符号秩统计量 $W^+ = \sum S_i R_i^+ = 28$，查附表 3 得其检验的 p 值为 $P_{H_0}(W^+ \geqslant 28) = 0.0977$，如取 $\alpha = 0.05$，则我们不能拒绝原假设.

对于成对数据，我们仍然可以用 Wilcoxon 符号秩检验进行，望读者自己考虑.

当总体分布函数并非处处连续时，样本中可能有相等的出现，我们说存在着结. 例如，设有 4 个样本，其依次为：1.3，1.7，1.7，2.5 把它们由小到大排序之后知道，$R_1 = 1$，$R_4 = 4$，而 (R_2, R_3) 可能取(2,3)也能取(3,2)，这样就有一个取法问题. 一般地，有两种方法处理该问题：一是随机化法；二是平均秩法(midrank)(还有其它方法，详见[9]). 所谓随机化法，即是对同一结内的样本，按该结所占据的秩，用等概率的方法配秩. 对于上面的例子，有 $P((R_2,R_3) = (2,3)) = P((R_2,R_3) = (3,2)) = \frac{1}{2}$. 随机化方法的最大优点是定理 3.3，定理 3.4，定理 3.5，定理 3.6 的结论仍成立，这对讨论某些统计量的确切分布是有很大好处的，但是，由于额外地加入了一个随机化，导致它有一个致命的缺点：结果因人而异，不可重复. 我们看一个例子.

例 3.4 设有 9 个样本如下：

i	1	2	3	4	5	6	7	8	9
X_i	-5	7	4	3	5	-7	5	-6	5
S_i	0	1	1	1	1	0	1	0	1
R_i^+			2	1				7	

其中 $|X_1| = X_5 = X_7 = X_9 = 5$，其秩可取(3,4,5,6)中的任一排列；$X_2 = |X_6| = 7$，其秩可取(8,9)中的任一排列. 如果 Wilcoxon 符号秩检验的拒绝域为 $\{W^+ \geqslant 26\}$，则按随机化方法取秩，甲乙两人分别取秩如下：

i	1	2	3	4	5	6	7	8	9
甲	3	9	2	1	4	8	5	7	6
乙	6	8	2	1	3	9	4	7	5

甲计算得 $W^+ = 27 > 26$，拒绝 H_0；而乙计算得 $W^+ = 23 < 26$，不能拒绝 H_0. 这显然是不能接受的.

基于上面随机化方法定秩的缺点，我们一般都采用下面要讲的平均秩法. 为此我们引进一个结统计量的概念.

设样本 $X_1,\cdots,X_n$ 由小到大如下排列：

$$X_{(1)} = X_{(2)} = \cdots = X_{(\tau_1)}$$

$$< X_{(\tau_1 - 1)} = \cdots = X_{(\tau_1 + \tau_2)} < \cdots < X_{(\tau_1 + \cdots + \tau_{g-1} + 1)} = \cdots = X_{(\tau_1 + \cdots + \tau_g)}$$

其中 $(\tau_1, \cdots, \tau_g)$ 是 g 个正整数，且 $\sum_i \tau_i = n$，则称$(\tau_1, \cdots, \tau_g)$ 为结统计量.

对于上一定义，我们应注意到：

1° g 是样本中结的个数，为随机的；

2° τ_i 是第 i 个结的长度，为随机的；

3° 对于样本 $X_1, \cdots, X_n$，上述结统计量将其分成 g 个组，按平均秩方法定秩，其第 i 组的样本均取秩为

$$m_i = \frac{1}{\tau_i} \sum_{k=1}^{\tau_i} (\tau_1 + \cdots + \tau_{i-1} + k) = \tau_1 + \cdots + \tau_{i-1} + \frac{1 + \tau_i}{2}$$

于是 n 个样本只取 g 个不同的秩，是唯一的，避免了随机化方法的结果不可重复性. 但是这样得到的样本秩有可能是非整数，故前面讲的定理 3.3 等并不成立，这就导致其理论处理不容易，但为了结论的科学性，我们都是利用平均秩法.

下面我们以一个例子说明，当有结存在时，如何求 Wilcoxon 符号秩统计量的零分布.

例 3.5 设有 4 个数据如下：

X_i	1	2	−1	2
平均秩	4.5	6.5	4.5	6.5
符号秩	+4.5	+6.5	−4.5	+6.5

则 Wilcoxon 符号秩统计量 $W^+ = 4.5 + 6.5 + 6.5 = 17.5$. 由于其共有 $2^4 = 16$ 个不同的符号秩，故其零分布为

4.5	4.5	6.5	6.5	$W^+ = k$	$P(W^+ = k)$
−	−	−	−	0	1/16
+	−	−	−	4.5	2/16
−	+	−	−	4.5	2/16
−	−	+	−	6.5	2/16
−	−	−	+	6.5	2/16
+	+	−	−	9	1/16
+	−	−	−	11	4/16
+	−	−	+	11	4/16
−	+	+	−	11	4/16
−	+	−	+	11	4/16
−	−	+	+	13	1/16
+	+	+	−	15.5	2/16
+	+	−	+	15.5	2/16
−	+	+	+	17.5	2/16
+	−	+	+	17.5	2/16
+	+	+	+	22	1/16

则检验的 p 值为 $P_{H_0}(W^- \geqslant 17.5) = \frac{3}{16}$.

虽然对于有结的 Wilcoxon 符号秩统计量的零分布无表可查,但是当 n 很大时,它有如下的渐近正态性:

$$\frac{W^+ - E_{H_0}(W^+)}{\sqrt{\mathrm{Var}_{H_0}(W^+)}} \xrightarrow{\mathscr{L}} N(0,1)$$

其中

$$E_{H_0}(W^+) = \frac{n(n+1) - d_0(d_0+1)}{4}$$

$$\mathrm{Var}_{H_0}(W^+) = \frac{n(n+1)(2n+1) - d_0(d_0+1)(2d_0+1)}{24} - \frac{\sum_{i=1}^{g}(\tau_i^3 - \tau_i)}{48}$$

这里 $(\tau_1, \cdots, \tau_g)$ 为结统计量,d_0 是差值为零的个数. 此结论的证明见[14].

例 3.6 在研究维生素 B_1 对学习影响的过程中,从孤儿院中随机地抽取 74 名儿童,并随机地把他们分成 37 对,从每一对中随机地选取一个服用维生素 B_1,另一个服用一种无药效的安慰剂. 服用六周后,其中 12 对儿童的智商(IQ)值增加如下:

对	2	5	8	11	14	17	20	23	26	29	32	35
吃 V_1	14	18	2	4	−5	14	−3	−1	1	6	3	3
不吃	8	26	−7	−1	2	9	0	−4	13	3	3	4
差	6	−8	9	5	−7	5	−3	3	−12	3	0	−1
秩	8	−10	11	6.5	−9	6.5	−4	4	−12	4	0	2
符号	+	−	+	+	−	+	−	+	−	+		−

试问服用维生素 B_1 对提高智商是否有影响?

对上面的数据,其结统计量为

$$\tau_1 = 1, \tau_2 = 1, \tau_3 = 3, \tau_4 = 2, \tau_5 = \tau_6 = \cdots = \tau_9 = 1, d_0 = 1$$

又由于

$W^+ = 40$

$E(W^+) = \frac{1}{4}(12 \times 13 - 1 \times 2) = 38.5$

$\mathrm{Var}(W^+) = \frac{1}{24}(12 \times 13 \times 25 - 1 \times 2 \times 3) - \frac{1}{48}(2^3 - 2 + 3^3 - 3) = 161.625$

则检验的 p 值近似地等于

$$P(W^+ \geqslant 40) = 1 - \Phi\left(\frac{40 - 38.5}{\sqrt{161.625}}\right) = 0.453$$

故认为维生素 B_1 对 IQ 的提高无多大帮助.

3.4 点估计和区间估计

在上一节的基础上,我们要给出基于某些检验统计量的关于对称中心 θ 的点估计和区间

估计. 首先我们把样本 $X_1,\cdots,X_n$ 扩大一下. 最简单的扩大就是加入每两个数的平均，这就是所谓的 Walsh 平均(Walsh averages). 这 $\frac{n(n+1)}{2}$ 个平均是 $\frac{X_i+X_j}{2}$，$i\leqslant j$. 我们有以下很方便的定理：

定理 3.7　假设分布函数 $F(x)\in\Omega_0$，且 $X_1,\cdots,X_n\overset{\text{iid}}{\sim}F(x)$，则 Wilcoxon 符号秩统计量 W^+ 等于正的 Walsh 平均之数目：

$$W^+=\#\left\{\frac{X_i+X_j}{2}>0,i\leqslant j\right\}$$

证明　用 $X_{i_1},\cdots,X_{i_p}$ 表示正的样本点，画半开区间 $I_1=(-X_{i_1},X_{i_1}]$. 显然 X_{i_1} 的秩(按绝对值) $R_{i_1}^+$ 等于在 I_1 中的所有样本点数. 而 I_1 中所有样本点和 X_{i_1} 的 Walsh 平均都大于 0. 因此 $R_{i_1}^+$ 等于大于 0 的由 I_1 中样本点和 X_{i_1} 所形成的 Walsh 平均的数目. 一般地，我们可构造 I_{i_k}，$k=1,\cdots,p$，而且 $R_{i_k}^+$ 为大于 0 的由 I_k 中样本点和 X_{i_k} 所形成的 Walsh 平均的数目. 因为 W^+ 等于这些 $R_{i_k}^+$ 的和，所以也等于所有大于 0 的 Walsh 平均的数目.　□

上面的思路导致定义统计量

$$W^+(\theta)=\#\left\{\frac{X_i+X_j}{2}>\theta;i\leqslant j\right\}$$

用 $W^+(\theta_0)$ 作为检验 $H_0:\theta=\theta_0$ 对 $H_1:\theta>\theta_0$ 的统计量，则检验是无偏的.（无偏检验：在 H_0 下，拒绝零假设的概率不大于水平 α，而在 H_1 下，拒绝零假设的概率不小于 α.）

当样本 $X_1,\cdots,X_n\overset{\text{iid}}{\sim}F(x-\theta)$，$F(x)\in\Omega_S$ 时，我们用 Walsh 平均的中位数

$$\hat{\theta}=\text{median}\left\{\frac{X_i+X_j}{2},i\leqslant j=1,\cdots,n\right\}$$

作为 θ 的估计量，称为基于 Wilcoxon 符号秩统计量 W^+ 的关于 θ 的 Hodges-Lehmann 估计量(HL 估计量).

相应于更广泛的检验统计量，我们可定义一般的 Hodges-Lehmann 估计量. 假设 V 是一个检验统计量(零假设 $H_0:\theta=0$)，统计量 $V(\theta)$ 是把 V 中的 X_i 换成 $X_i-\theta$ 而得. 仍然假定 $V(\theta)$ 是 θ 的非增函数，而且在零假设下，$V=V(0)$ 对称于某 μ_0. 对于来自 $F(x)(F(x-\theta)\in\Omega_S)$ 的随机样本 $X_1,\cdots,X_n$，定义

$$\theta^*=\sup\{\theta:V(\theta)>\mu_0\},\qquad\theta^{**}=\inf\{\theta:V(\theta)<\mu_0\}$$

则

$$\hat{\theta}=\frac{\theta^*+\theta^{**}}{2}$$

称为 θ 的 Hodges-Lehmann 估计量. 以符号统计量为例，当 n 为偶数时 $\theta^*=X_{(\frac{n}{2})}$，$\theta^{**}=X_{(\frac{n}{2}+1)}$；当 n 为奇数时 $\theta^*=\theta^{**}=X_{(\frac{n+1}{2})}$. 因此，$\hat{\theta}=\text{median}\{X_i,i=1,\cdots,n\}$，即样本中位数；对 Wilcoxon 符号秩统计量 W^+，有 $\hat{\theta}=\text{median}\left\{\frac{X_i+X_j}{2},i\leqslant j=1,\cdots,n\right\}$，已在前面引进，特别地，当 $n=3$ 时，$\hat{\theta}=\frac{1}{4}(X_{(1)}+2X_{(2)}+X_{(3)})$，从此可以看出，Hodges-Lehmann 估计对某些样本有所侧重，同时又兼顾所有样本中的信息，对于传统的 t 统计量，$\hat{\theta}=\overline{X}$，即样本均值.

关于 Hodges-Lehmann 估计的一些分布性质，可参见习题 5，6，详细的请见[7]. 下面我们给出一个对称性的结论.

定理 3.8　如果总体分布 $F(x-\theta)\in\Omega_S$，则 Hodges-Lehmann 估计量 $\hat{\theta}$ 的分布关于 θ 对称.

证明　因为 $P_\theta(\hat{\theta}-\theta<x)=P_0(\hat{\theta}<x)$，我们只需考虑 $\theta=0$ 的情况. 这时 $X=(X_1,\cdots,X_n)$ 和 $-X=(-X_1,\cdots,-X_n)$ 同分布，因此，$\hat{\theta}(X)$ 和 $\hat{\theta}(-X)$ 同分布. 从 $\hat{\theta}$ 的定义，$\hat{\theta}(-X)=-\hat{\theta}(X)$，即 $\hat{\theta}$ 与 $-\hat{\theta}$ 同分布.　　□

下面我们再考虑一个更广义的统计量 $V(\theta)$ 的性质，它定义为

$$V(\theta)=\sum_{j=1}^{n}a(R_j^+(\theta))S(X_j-\theta)=\sum_{j=1}^{n}a_jS(X_j)=\sum_{j=1}^{n}a_jW_j$$

这里 $R_j^+(\theta)$ 是 $|X_j-\theta|$ 在 $|X_1-\theta|,\cdots,|X_n-\theta|$ 中的秩，其它符号是以前介绍过的. $V(\theta)$ 有下面的性质，该定理的证明直观性很强，请读者自己验证.

定理 3.9　$V(\theta)$ 是 θ 的非增阶梯函数，它在每个 $X_{(i)}$ 点下降 a_1，而在 $\frac{X_{(i)}+X_{(j)}}{2}$ $(j>i)$ 点下降 $a_{j-i+1}-a_{j-i}$.

注：1. 如果随机样本 $X_1,\cdots,X_n$ 来自 $F(x-\theta)$，$F(x)\in\Omega_S$，则 $V(\theta)$ 关于 $\sum_{i=1}^{n}\frac{a_i}{2}$ 对称.

2. Hodges-Lehmann 估计量也可被关系 $V(\hat{\theta})=\sum_{i=1}^{n}\frac{a_i}{2}$ 所决定.

用 Walsh 平均，还可得到区间估计. 令 $W_{(1)},\cdots,W_{(N)}$ 为升幂排列的 Walsh 平均($N=\frac{n(n+1)}{2}$)，而且

$$P(W^+<a)=\frac{\alpha}{2}=P(W^+\geqslant N-a)$$

则

$$[W_{(a+1)},W_{(N-a)})$$

为 $(1-\alpha)100\%$ 置信区间. 这里 a 既可用 W^+ 的分布算，也可以有

$$a\approx\frac{n(n+1)}{4}-0.5-Z_{\frac{\alpha}{2}}\sqrt{\frac{n(n+1)(2n+1)}{24}}$$

上面我们讲了有关 Hodges-Lehmann 估计的表达式及分布性质，下面则给出基于 Wilcoxon 符号秩统计量的 HL 估计

$$\hat{\theta}_W=\text{median}\left\{\frac{X_i+X_j}{2},i\leqslant j=1,\cdots,n\right\}$$

的手工计算方法，以一个例子说明.

例 3.7　设样本为：62,70,74,75,77,80,83,85,88. 下面我们用图示法求基于 Wilcoxon 符号秩统计量 W^+ 的关于 θ 的 HL 估计 $\hat{\theta}_W$，此时 $N=45$.

1°　如图 3.1，在一个直角坐标系上画直线 $y=x$；

2°　在直线 $y=x$ 上，点上样本点(X_i,X_i)，$i=1,\cdots,n$；

3°　在坐标系上标出(X_i,X_j)，$i\geqslant j$，则 Walsh 平均就为这些点相应的 x，y 坐标的平均值 $\frac{X_i+X_j}{2}$，$i\geqslant j$；

4°　作一条垂直于对角线 $y=x$ 的直线 l_1，使之从原点向上移动(或者作另一条垂直于

$y=x$ 的直线 l_2，使之从上向下移动），直至第 $\frac{N+1}{2}=23$ 个点落在直线 l_1（或者 l_2 上），此时直线 l_1 上有两点：(80,75)，(85,70)，任取其一的横、纵坐标的平均值即为 HL 估计

$$\hat{\theta}_W=\frac{80+75}{2}=77.5=\frac{85+70}{2}$$

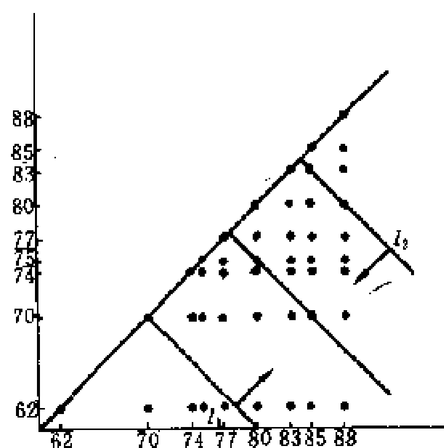


图 3.1　HL 估计的计算

当 N 为偶数时，则移动 l_1（或者 l_2），找其第 $\frac{N}{2}$ 与 $\frac{N}{2}+1$ 个点计算这两点的 Walsh 平均后，再求平均即可. 当有结存在时，在坐标系的点上注明结的长度即可.

前面我们仅介绍了在 $\theta=0$ 时 W^+ 的分布. 下面考虑随机样本 $X_1,\cdots,X_n$ 来自一个任意连续分布 $H(x)$. 令

$$p_1=P(X_1>0) \qquad p_2=P(X_1+X_2>0)$$

$$p_3=P(X_1-X_2>0,X_1>0) \qquad p_4=P(X_1+X_2>0,X_1+X_3>0)$$

定理 3.10　对于 Wilcoxon 符号秩统计量 W^+，

$$E(W^+)=np_1+\frac{n(n-1)}{2}p_2$$

$$\text{Var}(W^+)=np_1(1-p_1)-\frac{n(n-1)}{2}p_2$$

$$+2n(n-1)(p_3-p_1p_2)+n(n-1)(n-2)(p_4-p_2^2)$$

证明见[9]. 定理 3.10 在考虑检验的势及效率等问题时有用. 读者可验证前面在 $\theta=0$ 时的结果是这个结果的特殊情况.

3.5　渐近相对效率及比较

假设 $X_1,\cdots,X_n\overset{\text{iid}}{\sim}F(x-\theta)$，$F(x)\in\Omega_S$. 根据第一章，只要 Pitman 条件满足，我们可通过求 $\mu'_n(0)$ 和 $\sigma_n(0)$ 来找到一个统计量的效率 c，从而可用不同统计量的效率得到渐近相对效率

(ARE). 下面举几个例子. 我们用 $f(x)$ 表示 $F(x)$ 的概率密度函数.

1. 记符号统计量 $S = \#(X_i > 0, 1 \leqslant i \leqslant n)$，有

$$E(S) = n(1 - F(-\theta)), \qquad \mathrm{Var}(S) = n(1 - F(-\theta))F(-\theta)$$

可取 $\mu_n(\theta) = E(S)$ 及 $\sigma_n^2(\theta) = \mathrm{Var}(S)$，于是有

$$\mu'_n(0) = nf(0), \qquad \sigma_n^2(0) = \frac{n}{4}, \qquad c_S = 2f(0)$$

这里 c_S 表示符号统计量的效率.

2. 对 Wilcoxon 符号秩统计量 $W^+ = \sum_{j=1}^{n} R_j S(X_j)$，有

$$E(W^+) = np_1 + n\frac{n(n-1)}{2}p_2, \qquad \mathrm{Var}(W^+) = \frac{n(n+1)(2n+1)}{24}$$

可取 $\sigma_n^2(\theta) = \mathrm{Var}(W^+)$ 及

$$\mu_n(\theta) = E(W^+) = n(1 - F(-\theta)) + \frac{n(n-1)}{2}\int(1 - F(-x-\theta))f(x-\theta)\mathrm{d}x$$

有
$$\mu'_n(0) = nf(0) - n(n-1)\int f^2(x)\mathrm{d}x, \qquad c_{W^+} = \sqrt{12}\int f^2(x)\mathrm{d}x$$

这里 c_{W^+} 表示 Wilcoxon 符号秩统计量的效率.

3. 对传统的 t 统计量，记 $\sigma_f = \int x^2 f(x)\mathrm{d}x$. 取

$$\mu_n(\theta) = \sqrt{n}\,\frac{\theta}{\sigma_f}, \qquad \sigma_n(0) = 1$$

有 $c_t = \frac{1}{\sigma_f}$. 这里 c_t 表示 t 统计量的效率.

由 ARE 的定义，$e_{12} = \frac{c_1^2}{c_2^2}$，则我们有上述三个统计量之间的 ARE：

$$\mathrm{ARE}(S, W^+) = \frac{c_S^2}{c_{W^+}^2} = \frac{f^2(0)}{3\left(\int f^2(x)\mathrm{d}x\right)^2}$$

$$\mathrm{ARE}(S, t) = \frac{c_S^2}{c_t^2} = 4\sigma_f^2 f^2(0)$$

$$\mathrm{ARE}(W^+, t) = \frac{c_{W^+}^2}{c_t^2} = 12\sigma_f^2\left(\int f^2(x)\mathrm{d}x\right)^2$$

因此，对任意给定的分布，我们都可计算上面的 ARE，见下表：

分布	$U(-1,1)$	$N(0,1)$	logistic	重指数
密度	$\frac{1}{2}I(-1,1)$	$\frac{\exp\left(-\frac{x^2}{2}\right)}{\sqrt{2\pi}}$	$e^{-x}(1+e^{-x})^{-2}$	$\frac{e^{-\lvert x\rvert}}{2}$
$\mathrm{ARE}(W_n^+, T_n; F)$	1	$\frac{3}{\pi}$	$\frac{\pi^2}{9}$	$\frac{3}{2}$
$\mathrm{ARE}(S_n^+, T_n; F)$	$\frac{1}{3}$	$\frac{2}{\pi}$	$\frac{\pi^2}{12}$	2

下面例子讨论了正态分布有不同程度"污染"时，ARE(W^+,t) 的不同结果.

例 3.8　假定随机样本$X_1,\cdots,X_n$来自$F_\varepsilon=(1-\varepsilon)\Phi(x)+\varepsilon\Phi(\frac{x}{3})$. 这里$\Phi(x)$为$N(0,1)$的分布函数. 易见，

$$\int f_\varepsilon^2(x)\mathrm{d}x=\frac{(1-\varepsilon)^2}{2\sqrt{\pi}}+\frac{\varepsilon^2}{6\sqrt{\pi}}+\frac{\varepsilon(1-\varepsilon)}{\sqrt{5\pi}},\qquad \sigma_{F_\varepsilon}^2=1+8\varepsilon$$

由上面公式得

$$\mathrm{ARE}(W^+,t)=\frac{3(1+8\varepsilon)}{\pi}\left[(1-\varepsilon)^2+\frac{\varepsilon^2}{3}+\frac{2\varepsilon(1-\varepsilon)}{\sqrt{5}}\right]^2$$

对不同的 ε，我们有下表：

ε	0	0.01	0.03	0.05	0.08	0.10	0.15
ARE(W^+,t)	0.955	1.009	1.108	1.196	1.301	1.373	1.497

从上面两个表可以看出，只用到样本中大小次序方面信息的 Wilcoxon 符号秩检验、符号检验和 t 检验最具有优势的情况，即 F 为 $N(0,1)$ 时相比，效率并不算差. 对于其它几个 t 检验不占优势的情况，即 F 不为正态时，W_n^+ 基本上都优于 t 检验. 但在总体分布偏离正态时，偏离越多，Wilcoxon 符号秩检验就越好. 可以证明，对任何总体分布，Wilcoxon 符号秩检验对 t 检验的渐近相对效率绝不少于 0.864，详见阅读知识一节.

我们以前说过，一个检验统计量及与其相联的估计量有同样的效率. 上面的符号统计量、Wilcoxon 符号秩统计量和 t 统计量分别相应于样本中位数、Walsh 平均的中位数及样本均值. 这些都是 Hodges-Lehmann 估计量的特例. 一般地有下面的估计效率 c 的定理：

定理 3.11　假设 $\hat{\theta}$ 为相应于满足 Pitman 条件的统计量 V 的 Hodges-Lehmann 估计量. 如果 V 的效率为 c，则

$$\lim_{n\to\infty}P(\sqrt{n}(\hat{\theta}-\theta)<a)=\Phi(ac)$$

即渐近地有 $\sqrt{n}(\hat{\theta}-\theta)\sim N(0,c^{-2})$.

证明见[14].

下表为 t 检验(t)，符号检验(S)，Wilcoxon 符号秩检验(W^+) 之间的 ARE 的范围，其中带星号(*)的为分布是非单峰时的结果.

	t	S	W^+
t	—	$(0,3];(0,\infty)^*$	$\left(0,\frac{125}{108}\right)$
S	$[\frac{1}{3},\infty);(0,\infty)^*$	—	$[\frac{1}{3},\infty);(0,\infty)^*$
W^+	$\left(\frac{108}{125},\infty\right)$	$(0,3];(0,\infty)^*$	—

例如，由上表可看出 $0.864 = \frac{108}{125} < \mathrm{ARE}(W^+, t) < \infty$，无穷是在 Cauchy 分布时出现. 很明显，在分布未知时，非参数方法有很大的优越性. 在用 Pitman 渐近相对效率时，要注意这个概念只对大的样本适用，并且它只局限在 H_0 点的一个邻域中比较.

3.6 阅读知识

3.6.1 符号秩的一般分布

前面的定理 3.4—3.6，叙述了符号秩的确切分布，但是它们还只是给出了绝对值样本秩的零分布，而下一定理对符号秩的分布刻画更加明了.

定理 3.12 设 $X_1, \cdots, X_n \overset{iid}{\sim} F(x) \in \Omega_0$，又设 $Q = \sum_i I(X_i > 0)$，是随机的，如果 $Q = q$，则 $S_1 < \cdots < S_q$ 表示 $X_1, \cdots, X_n$ 的符号秩由小到大的排序，对于 $q = 0, 1, \cdots, n$ 有

$$P(Q = q, S_1 = s_1, \cdots, S_q = s_q) = \begin{cases} \frac{1}{2^n}, & 1 \leqslant s_1, < \cdots < s_q \leqslant n \\ 0, & \text{否则} \end{cases}$$

此定理的证明并不难，读者自己作为练习试一试或见[7].

3.6.2 Wilcoxon 符号秩统计量的极限分布的证明

关于 wilcoxon 符号秩统计量的极限分布的证明，可用第八章讲的一般秩分布理论. 下面我们先用其与 U 统计量的关系给出证明.

我们回忆一下例 1.76 中的单样本 U 统计量可以表示为(H_0 下)：

$$U_n = \frac{1}{\binom{n}{2}}\left(\sum_i R_i^+ - r\right)$$

现在看来 R_i^+ 就是本章讲的符号秩，$r = \sum_i I(X_i > 0) \equiv S_n^+$ 就是上章的符号秩统计量. 即在 H_0 下

$$W_n^+ = \binom{n}{2} U_n + S_n^-$$

又因为在 H_0 下

$$E_{H_0} U_n = P(X_1 + X_2 > 0) = \frac{1}{2} \qquad E_{H_0} S_n^+ = \frac{n}{2}$$

则在 H_0 下，由定理 1.7 知道

$$S_n^- - \frac{n}{2} \xrightarrow{a.s.} 0$$

$$\sqrt{n}\left(U_n - \frac{1}{2}\right) \xrightarrow{\mathscr{L}} N\left(0, \frac{1}{3}\right)$$

又由于在 H_0 下

$$\frac{\sqrt{n}\left(W_n^+ - \frac{n(n+1)}{4}\right)}{\binom{n}{2}} = \sqrt{n}\left(U_n - \frac{1}{2}\right) + \frac{\sqrt{n}\left(S_n^+ - \frac{n}{2}\right)}{\binom{n}{2}}$$

且 $\frac{\sqrt{n}}{\binom{n}{2}} \longrightarrow 0$，则 $\frac{\sqrt{n}\left(W_n^+ - \frac{n(n+1)}{4}\right)}{\binom{n}{2}}$ 与 $\sqrt{n}\left(U_n - \frac{1}{2}\right)$ 同分布，所以

$$\frac{\sqrt{3n}}{\binom{n}{2}}\left(W_n^+ - \frac{n(n+1)}{4}\right) \xrightarrow{\mathscr{L}} N(0,1)$$

又由于

$$\frac{\sqrt{\frac{n(n+1)(2n+1)}{24}}}{\frac{\binom{n}{2}}{\sqrt{3n}}} \longrightarrow 1$$

则

$$\frac{W_n^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \xrightarrow{\mathscr{L}} N(0,1)$$

3.6.3 $\mathrm{ARE}(W_n^+, t; F) \geqslant 0.864$ 的证明

在§3.5中我们指出

$$\mathrm{ARE}(W_n^-, t; F) \geqslant 0.864$$

这是Hodges-Lehmann于1955年用数学分析方法证明的，我们简述于下．首先假设总体分布函数满足条件

1° $F(x)$ 有概率密度函数 $f(x)$ 且 $f(x)$ 在原点连续，$f(0) > 0$，并且 $\int x^2 f(x)\mathrm{d}x$ 与 $\int f^2(x)\mathrm{d}x$ 均存在、有限；

2° $F(x)$ 关于原点对称．

不妨设总体方差 $\sigma_f^2 = 1$，则要证明

$$12\left(\int f^2(x)\mathrm{d}x\right)^2 \geqslant 0.864$$

事实上，取 $f_0(x) = \frac{3}{20\sqrt{5}}(5 - x^2)I(|x| < \sqrt{5})$，则

$$\begin{aligned}\int f^2(x)\mathrm{d}x &= \int (f(x) - f_0(x) + f_0(x))^2 \mathrm{d}x \\ &= \int (f(x) - f_0(x))^2 \mathrm{d}x + \int f_0^2(x)\mathrm{d}x + 2\int f_0(x)(f(x) - f_0(x))\mathrm{d}x\end{aligned}$$

又因为

$$\int f_0^2(x)\mathrm{d}x = \frac{3}{5\sqrt{5}}$$

$$\int f_0(x)f(x)\mathrm{d}x = \int_{-\sqrt{5}}^{\sqrt{5}} \frac{3}{20\sqrt{5}}(5-x^2)f(x)\mathrm{d}x$$

$$\geqslant \int_{-\infty}^{\infty} \frac{3}{20\sqrt{5}}(5-x^2)f(x)\mathrm{d}x = \frac{3}{5\sqrt{5}} = \int_{-\infty}^{\infty} f_0^2(x)\mathrm{d}x$$

则

$$\int_{-\infty}^{\infty} f_0(x)(f(x)-f_0(x))\mathrm{d}x \geqslant 0$$

所以

$$\int_{-\infty}^{\infty} f^2(x)\mathrm{d}x \geqslant \int_{-\infty}^{\infty}(f(x)-f_0(x))^2\mathrm{d}x + \frac{3}{5\sqrt{5}} \geqslant \frac{3}{5\sqrt{5}}$$

即

$$12\left(\int_{-\infty}^{\infty} f^2(x)\mathrm{d}x\right)^2 \geqslant 12\cdot\frac{9}{25\cdot 5} = 0.864$$

3.7 习 题

1. 证明推论 3.1.
2. 证明推论 3.2.
3. 对于密度函数

$$f(x) = \begin{cases} \dfrac{3}{20\sqrt{5}}(5-x^2), & |x| < \sqrt{5} \\ 0, & \text{否则} \end{cases}$$

证明

$$\mathrm{ARE}(W_n^+, T_n, F) = 0.864$$

(这说明 0.864 的下界能达到).

4. 证明 Willcoxon 符号秩统计量 W_n^+ 的零分布关于其期望对称.

5. 设 $\hat{\theta}(X_1,\cdots,X_n)$ 为基于检验统计量 $V(X_1,\cdots,X_n)$ 的关于 θ 的 HL 估计(此统计量满足所需条件),证明对于任意的 k 和 $X_1,\cdots,X_n$

$$\hat{\theta}(X_1+k,\cdots,X_n+k) = k + \hat{\theta}(X_1,\cdots,X_n)$$

6. 设 $X_1,\cdots,X_n \overset{\text{iid}}{\sim} F(x)$,分布函数 $F(x)$ 连续且关于原点对称,$V(X_1,\cdots,X_n)$ 为满足 HL 估计条件的检验统计量,$\hat{\theta}(X_1,\cdots,X_n)$ 为基于 $V(X_1,\cdots,X_n)$ 关于 θ 的 HL 估计. 如对于任何的 $X_1,\cdots,X_n$,$V(X_1,\cdots,X_n)+V(-X_1,\cdots,-X_n)=2\xi$,则 $\hat{\theta}(X_1,\cdots,X_n)$ 是一个奇统计量且关于 θ 对称.

7. 设 X 为一个随机变量,我们称 X 是关于某点 c 加权对称的,如果存在常数 λ,使

$$P(X > c+z) = \lambda P(X < c-z), \quad \forall z > 0$$

(1)如果 X 关于某常数 c 加权对称,则在 $P(X<c)>0$ 下,有

$$\lambda = \frac{P(X>c)}{P(X<c)}$$

(2)证明在 $0<P(X<c)=P(X\leqslant c)<1$ 条件下,随机变量 $|X-c|$ 和 $S_c = I(X>c)$ 是独立的.

8. 设 W^+ 为 Willcoxon 符号秩统计量,证明在 $H_0:\theta=0$ 下,

$$W^+ \overset{d}{=} \sum_{j=1}^{n} V_j$$

其中 $V_1, \cdots, V_n$ 是相互独立的且 $P(V_j = j) = P(V_j = 0) = \dfrac{1}{2}, j = 1, \cdots, n$.

9. 对于单样本对称中心检验问题，如假设总体分布连续且记 $W^- = \sum_{i=1}^{n} R_i^+ I(X_i < 0)$.

(1)证明 $W^+ = \dfrac{n(n+1)}{2} - W^-$；

(2)设 $W_0 = W^+ - W^-$，则证明在 $H_0: \theta = 0$ 下，W_0 是关于 0 对称的.

10. 设 $X_1, \cdots, X_n \overset{\text{iid}}{\sim} F(x)$，$F(x)$ 连续，以 R_i 记 X_i 在 $X_1, \cdots, X_n$ 中的秩，设 $n \geqslant 2$，并且记 $V = R_1 - R_n$，证明

$$P(V = k) = \begin{cases} \dfrac{n - |k|}{n(n-1)}, & |k| = 1, \cdots, n-1 \\ 0, & \text{否则} \end{cases}$$

11. 设 $Z_1, \cdots, Z_N$ 为来自某连续分布函数 $F(x)$ 的 iid 样本，且 $F(x)$ 关于原点对称. 令 $m = \sum I(Z_i < 0)$，$n = \sum I(Z_i > 0)$，设 $X_1, \cdots, X_m$ 和 $Y_1, \cdots, Y_n$ 为 Z_i 中小于零与大于零的绝对值的样本，试回答：基于 Z 样本的 Wilcoxon 秩统计量 W^+ 与基于 X, Y 样本的 Wilcoxon 秩统计量 W 有何关系？

12. 设 W^+ 为关于 $H_0: \theta = \theta_0$ 的 Wilcoxon 符号秩统计量(样本容量为 n).

(1)证明，W^+ 的零分布为

$$P_{H_0}(W^+ = k) = \begin{cases} \dfrac{c_n(k)}{2^n}, & k = 0, 1, \cdots, \dfrac{n(n+1)}{2} \\ 0, & \text{否则} \end{cases}$$

其中 $c_n(k)$ 为 $\{1, \cdots, n\}$ 中子集元素之和等于 k 的子集个数；

(2)证明(1)中的 $c_n(k)$ 满足：$\forall\, k = 0, 1, \cdots, \dfrac{n(n+1)}{2}$ 及 $n \geqslant 2$

$$c_n(k) = c_{n-1}(k - n) + c_{n-1}(k)$$

其中

$$c_0(k) = \begin{cases} 1, & k = 0 \\ 0, & k \neq 0 \end{cases}$$

(3)利用(2)中的结果，给出 W^+ 的零分布的递推公式；

(4)对于 $n = 1, 2, 3, 4$，试用(3)的公式求 W^+ 的零分布.

13. 利用第 8 题的结论，验证 Wilcoxon 符号秩统计量的零分布的期望与方差.

14. 利用 Wilcoxon 符号秩统计量与 Walsh 平均的关系，对任意的对称中心 θ，求出 W^+ 的期望与方差.

15. 设 $X_1, \cdots, X_n \overset{\text{iid}}{\sim} F(x)$，$F(x)$ 连续，且关于原点对称，又设 $X_1, \cdots, X_n$ 中无零存在，证明 Wilcoxon 符号秩统计量 W^+ 的分布为

$$P(W^+ = k) = \sum_{j=0}^{n} \binom{n}{j} 2^{-n} P(W^+ = k \mid S = j), \quad k = 0, 1, \cdots, \frac{n(n+1)}{2}$$

其中 $S = \sum_{i=1}^{n} I(X_i > 0)$.

16. 设 $X_{(1)} \leqslant \cdots \leqslant X_{(n)}$ 为 n 个样本的顺序统计量，令 $W_{ij} = \dfrac{1}{2}(X_{(i)} + X_{(j)}), i \leqslant j$，则

(1) 上述 Walsh 平均中前三个最小的为 W_{11}, W_{12} 和 $\min\{W_{22}, W_{13}\}$，那么第四个最小的是什么？

(2)试证明，W_{ij} 在 Walsh 平均中最小可能的秩是 $\dfrac{i}{2}(2j - i + 1)$.

(3)证明，W_{1j} 在 Walsh 平均中最大可能的秩是 $\dfrac{1}{2}j(j-1) + 1$，且说明对于 $i \geqslant 2$，W_{ij} 最大可能的秩是

多少？

17. 设 $X_1,\cdots,X_n$ 为来自连续分布 $F(x)$ 的 iid 样本，且 $F(x)$ 关于 μ 对称. 证明：

(1)当 μ 增大时，$X_i-\mu$ 的符号秩仅当 μ 等于某个 Walsh 平均 $\frac{1}{2}(X_i+X_j)$ 时，才改变；

(2)对于(1)中的情形，当 μ 等于某个 Walsh 平均 $\frac{1}{2}(X_i+X_j)$ 时，如 $i=j$，则 X_i 的符号秩由 1 变到 -1；如 $i\neq j$，且 $X_i<X_j$，则 X_i 的符号秩由 $-(k+1)$ 变到 $-(k+2)$，其中 k 为 X_i 与 X_j 间的样本数.

18. 有人在研究得克萨斯州立大学篮球队员的体重时，记录了如下 15 名队员的体重数据：

188.0　211.2　170.8　212.4　156.9　223.1　235.9　183.9
214.4　221.0　162.0　220.8　174.1　210.3　195.2

试用 Wilcoxon 符号秩统计量检验其平均体重是否为 163.5，并写出其 p 值.

19. 在某小学，有人对刚入学的一年级 20 名学生进行了阅读测验，其得分如下：

33　19　40　51　41　27　23　39　21　37
41　31　46　51　34　37　36　55　52　32

由这些数据，我们能否说，入学新生的阅读分数小于 45？

20. 现比较某种产品的推销活动. 如果已知一个公司平均每月推销时间为 119(小时)，而现在从另一个公司的推销员中随机地抽取 16 名推销员，测其推销时间如下：

136　103　91　122　96　145　140　138
126　120　99　125　91　142　119　137

试问这一公司的推销员月均推销时间是否也为 119(小时)？

21. 某公司经理相信其雇员手工的灵活性得分大于 70. 现从该公司的雇员中随机地抽取 18 人进行检验，其结果如下：

94　91　84　80　58　46　47　49　76
86　87　93　65　48　85　59　72　71

试利用这些数据中的信息，以检验经理想法的可信度.

22. 一个食物研究所在检测某种香肠的肉含量时，随机地测得如下数据(%)：

76.5　74.1　73.8　80.4　77.8　76.9　68.3

(1)计算 Walsh 平均；

(2)求该种香肠肉含量的 HL 估计；

(3)求该种香肠肉含量的 95% 的基于 Wilcoxon 符号秩统计量的置信区间.

23. 某医院对病人要作某种手术需等待的天数感兴趣，现在随机地抽取 8 名病人，询问其所等待的天数如下：

6　11　15　9　12　7　13　9

(1)计算 Walsh 平均；

(2)求病人所等待天数的平均值的 HL 估计；

(3)求病人所等待平均天数的基于 Wilcoxon 符号秩统计量的 90%的置信区间.

24. 现随机地抽取 10 名 8 岁女孩的体重数据如下：

48 63 59 41 60 47 57 61 67 57

(1)试求 8 岁女孩的平均体重的 HL 估计；

(2)试求 8 岁女孩平均体重的基于 Wilcoxon 符号秩统计量的 90%的置信区间；

(3)试检验 8 岁女孩平均体重是否为 56.

25. 在研究交通监控系统时，先进行模拟研究，人们需要在模拟系统中的交通反馈时间为 60(秒). 现记录了 12 次反馈时间如下：

67 63 73 80 66 65 70 55 60 69 56 64

指出 H_0 与 H_1，并用 Wilcoxon 符号秩统计量检验之.

26. 现随机地抽取 36 名美国成年人，测其胆固醇含量如下：

251	145	260	257	243	289	204	168	186
234	321	244	458	299	269	217	175	220
98	303	212	248	224	326	289	233	196
289	250	256	266	265	275	252	222	249

(1)专家认为美国成年人的胆固醇含量是 210，试问这种论断是否合理?并指出 H_0，H_1(用 Wilcoxon 符号秩统计量)；

(2)给出基于 Wilcoxon 符号秩统计量的胆固醇含量的置信区间，并由此与(1)结论相比较.

27. 验证如下的 ARE：

分布	$U(-1,1)$	$N(0,1)$	logistic	重指数分布
ARE(S,t)	$\frac{1}{3}$	$\frac{2}{\pi}$	$\frac{\pi^2}{12}$	2

其中 S 为符号秩统计量，t 为传统的 t 检验统计量.

第四章　两样本问题

4.1 引　　言

实际问题中经常出现来自两个总体的样本之间的比较. 例如，比较两个班级某一门课程的成绩；比较两个不同品种小麦的产量；比较两个工艺的优劣；比较两种药物的效果等等. 传统上，人们假设总体是正态分布或近似的正态分布，然后利用两样本的 t 检验. 但是关于总体是正态的假设并不一定合理. 在小样本时，近似也不一定合适. 这时，如果用 t 检验，就可能犯错误. 事实上，这是个很常见的错误. 前面也提到过成对数据的比较问题，但那里的每个 X_i 只和 Y_i 一个数比较，与这里的整个 X 样本和整个 Y 样本比较不同. 在成对数据中，对每一样本，都受两个处理效应的影响，而我们感兴趣的却是其中的一个. 然而在两样本问题中，对每一样本它只受到一个处理效应的影响，恰好这就是我们感兴趣的. 由此看来二者是有着本质的不同.

在非参数统计中，我们对总体分布并不作什么假设，因此是解决这一类问题的好办法.

下面的例子是比较两个城市的高层建筑的高度. 城市 A 取了 9 个数据($X_1,\cdots,X_9$)，城市 B 取了 7 个数据($Y_1,\cdots,Y_7$)，列在下表中：

A	398	388	379	371	355	348	340	339	320
B	500	493	440	425	393	362	321		

这里没有前面讲的那样的天然的数据对. 事实上，这两个样本的大小也不一样. 但我们可以考虑比较所有可能的数据对. 我们也可以把两个样本混合并求其中位数，再把所有的样本点按样本及其相对于中位数的位置分类，按列联表的方法处理. 在应用各种方法之前，首先要对数据进行预先的分析，看其是否符合这些方法所要求的一些条件. 这些预先分析通常包括分别的或背靠背的茎叶图及盒子图等.

对于两样本问题，具体地讲是这样定义的：设

$$X_1,\cdots,X_m \overset{\text{iid}}{\sim} F_1\left(\frac{x-\theta_1}{\sigma_1}\right),\quad Y_1,\cdots,Y_n \overset{\text{iid}}{\sim} F_2\left(\frac{x-\theta_2}{\sigma_2}\right)$$

且 $X_1,\cdots,X_m,Y_1,\cdots,Y_n$ 相互独立，其中 θ,σ 为位置参数与刻度参数. 有关 θ_1 与 θ_2 的估计及假设检验，称为两样本位置参数问题；而有关 σ_1 与 σ_2 的估计和假设检验问题，称为两样本刻度参数问题，这两者统称为两样本问题. 由于以后常用的是两样本位置参数问题，故以后所谓的两样本问题，若非专门指出刻度参数，都是位置参数的，即如下的模型：

$$X_1,\cdots,X_m \overset{\text{iid}}{\sim} F(x),\quad Y_1,\cdots,Y_n \overset{\text{iid}}{\sim} F(x-\theta)$$

且 $X_1,\cdots,X_m,Y_1,\cdots,Y_n$ 是独立的. 参数 θ 称为位置参数,并且如果 $\theta>0$,则 Y 的分布向右平移,即 Y 倾向于比 X 来得大. 事实上

$$P(X>Y)=\int_{-\infty}^{\infty}\int_{-\infty}^{x}\mathrm{d}F(y-\theta)\mathrm{d}F(x)=\int_{-\infty}^{\infty}F(x-\theta)\mathrm{d}F(x)$$

则

$$P(X>Y)\leqslant\int_{-\infty}^{\infty}F(x)\mathrm{d}F(x)=\frac{1}{2}$$

基于上面 X,Y 样本的特性,可以得到许多的有关 $H_0:\theta=0$ 的秩检验方法. 下面我们将一一介绍.

4.2 中位数检验及 2×2 列联表

这里,$X_1,\cdots,X_m$ 及 $Y_1,\cdots,Y_n$ 为两个独立的随机样本. 它们来自两个有连续分布的总体,分别有未知的中位数 M_X 及 M_Y. 我们的目的是检验它们是否相同,即检验 $H_0:M_X=M_Y\leftrightarrow H_1:M_X\neq M_Y$(这里我们不考虑单边检验). 如果它们有相同的中位数 M_{XY},则必有 $P_X\equiv P(X>M_{XY})=P(Y>M_{XY})\equiv P_Y$,这里 X 和 Y 分别表示两个总体中的一般成员. 当然,作为第一步,我们先找出混合的样本中位数 M_{XY},然后再把所有样本点按其在 M_{XY} 的哪一边及来自哪个总体分成 4 部分. 这就形成了下面的列联表:

	X	Y	
$>M_{XY}$	A	B	$A+B$
$<M_{XY}$	C	D	$C+D$
	m	n	$N=m+n=A+B+C+D$

这里 A,B,C,D 分别为属于上述四个范畴的样本点数. 由初等概率论知,A,B 的联合分布为超几何分布:

$$P_{H_0}(A=k,B=l)=\frac{\binom{m}{k}\binom{n}{l}}{\binom{N}{k+l}}\qquad k=0,1,\cdots,m,\quad l=0,1,\cdots,n$$

但 A 和 B 本身分别为二项分布 $\mathrm{B}(m,p_X)$ 和 $\mathrm{B}(n,p_Y)$. 同样 $A+B$ 为 $\mathrm{B}(N,p)$.

当 H_0 成立时,A 与 C 应接近$\frac{m}{2}$,B 与 D 应接近$\frac{n}{2}$. 如果 H_1 为真,则$\left|\frac{A}{m}-\frac{B}{n}\right|$应倾向于取大值. 于是我们可以取检验统计量为$\left|\frac{A}{m}-\frac{B}{n}\right|$,而 H_0 的拒绝域应为其取大值.

由于超几何分布可用正态近似,且可用 $\hat{p}_X=\frac{A}{m}$,$\hat{p}_Y=\frac{B}{n}$ 及 $\hat{p}=\frac{A+B}{N}$ 来估计 p_X,p_Y 及 p,则对于大样本,有正态近似

$$Z = \frac{\hat{p}_X - \hat{p}_Y}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{m}+\frac{1}{n}\right)}} \xrightarrow{\mathscr{L}} N(0,1)$$

对于 § 4.1 中所给例子的数据，因为 $M_{XY} = 375$，则得到下面列联表：

	X	Y	
$>M_{XY}$	3	5	8
$<M_{XY}$	6	2	8
	9	7	16

我们有

$$\hat{p}_X = \frac{1}{3}, \quad \hat{p}_Y = \frac{5}{7}, \quad \hat{p} = \frac{1}{2}$$

因此得到 $Z = -1.5119$. 查附表2得该双边检验的 p 值为 0.1310，所以没有充分理由拒绝零假设.

注：如果发生某数据和 M_{XY} 相等，则删去此值，这时的 N 也要相应变化.

关于中位数检验统计量的取法，有的著作中取

$$M = \#\{Y_i > M_{XY}; i = 1,\cdots,n\}$$

可以证明，其零分布为

$$P_{H_0}(M = k) = \frac{\binom{n}{k}\binom{m}{\left[\frac{N}{2}\right]-k}}{\binom{N}{\left[\frac{N}{2}\right]}}, \quad k = 0,1,\cdots,n$$

感兴趣的读者可参见[7].

以后我们还要讨论对列联表数据的其它应用，同时还可以看到中位数检验对于厚尾的对称分布，是一个非常有效的检验(见下一节).

4.3 Mann-Whitney 检验

本节比较两个样本的方法是基于比较所有两个样本的可能的数据对，这等同于把两个样本混合排序，并比较两个样本的秩的大小. 这里并不假设总体分布的对称性，但要假设两个总体分布有类似的形状，这可由数据的预分析来验证. 有时可变换数据以获得所需的分布形状.

假定随机样本 $X_1,\cdots,X_m$ 和 $Y_1,\cdots,Y_n$ 分别来自 $F(x-M_X)$ 及 $F(x-M_Y)$(这表明两分布形状类似)，这里 $F(x)$ 为未知的分布函数. 令 $\theta = M_Y - M_X$. 不失一般性，我们要检验 H_0: $\theta = 0 \leftrightarrow H_1: \theta > 0$.

首先，把两个样本混合排序. 如用 R_i 表示第 i 个 Y 观察值 Y_i 在混合样本中的秩. 为使用符号方便，用 I_m 和 I_n 分别表示两样本的指标集. 令

$$R_i = \#(X_j < Y_i, j \in I_m) - \#(Y_k \leqslant Y_i, k \in I_n)$$

当 H_0 成立时，X,Y 样本为独立同分布的，而当 H_1 为真时，由 §4.1 知，$P(X>Y)<P(X<Y)$，这就是说 Y 样本倾向于大于 X 样本，即诸 R_i 倾向于取 $1,\cdots,N$ 中的后 n 个值. 于是 Wilcoxon 于 1945 年提出检验统计量 $W_Y = \sum_{i=1}^{n} R_i$，即 Y 样本的秩和，故称之为 Wilcoxon 秩和检验统计量. 显然，当 W_Y 很大时，应拒绝零假设.

又由例 1.7，我们可以看到

$$W_Y = \sum_{i=1}^{n} R_i = \#(X_j < Y_i, j \in I_m, i \in I_n) + \frac{n(n+1)}{2}$$

记 $W_{XY} = \#(X_j < Y_i, j \in I_m, i \in I_n)$，有 $W_Y = W_{XY} + \frac{n(n+1)}{2}$. 类似地，可定义 $W_{YX} = \#(X_j > Y_i, j \in I_m, i \in I_n)$ 及 $W_X = W_{YX} + \frac{m(m+1)}{2}$. 于是，$W_{XY} - W_{YX} = nm$. 显然，在零假设下，$W_{XY}$ 和 W_{YX} 同分布. W_Y 一般称为 Wilcoxon 秩和统计量(Wilcoxon rank-sum statistics)，而 W_{XY} 称为 Mann-Whitney 统计量(Mann-Whitney statistics). 因为这两个统计量对检验来说是等价的，和它们相关的检验也叫作 Wilcoxon 检验或 Mann-Whitney 检验. 为了解 W_Y(或 W_{XY})的分布性质，我们有下面简单的关于 R_i 的定理(留给读者证明).

定理 4.1　记 $N = n + m$. 在零假设下，对 $i \neq j$ 有

$$P(R_i = k) = \frac{1}{N}, \qquad k = 1,\cdots,N$$

$$P(R_i = k, R_j = l) = \begin{cases} \dfrac{1}{N(N-1)}, & k \neq l \\ 0, & k = l \end{cases}$$

容易验证

$$E(R_i) = \frac{N+1}{2}, \qquad \mathrm{Var}(R_i) = \frac{N^2-1}{12}, \qquad \mathrm{Cov}(R_i, R_j) = -\frac{N+1}{12}, \quad (i \neq j)$$

因为 $W_Y = \sum_{i=1}^{n} R_i$，则有

$$E(W_Y) = \frac{n(N+1)}{2}, \qquad \mathrm{Var}(W_Y) = \frac{mn(N+1)}{12}$$

显然

$$E(W_{XY}) = E(W_Y) - \frac{n(n+1)}{2} = \frac{mn}{2}$$

$$\mathrm{Var}(W_{XY}) = \mathrm{Var}(W_Y) = \frac{mn(N+1)}{12}$$

下面的例子表明如何对简单情况直接算出 W_{XY} 及 W_Y 的值和有关的概率，即零分布.

例 4.1　在 $m=n=2$ 时，所有可能的混合样本的秩为 1,2,3,4，相应于它们的数据来自 X 和 Y 样本的各种组合为 6 种. 在零假设下，每种可能的概率为 $p = \frac{1}{6}$. 下表为所有可能的组合及相应的 W_{XY} 及 W_Y 的值，左边的 1,2,3,4 为所有可能的秩.

1	Y	Y	Y	X	X	X
2	Y	X	X	Y	Y	X
3	X	Y	X	Y	X	Y
4	X	X	Y	X	Y	Y
W_{XY}	0	1	2	2	3	4
W_Y	3	4	5	5	6	7

由上表可得 W_{XY} 或 W_Y 在零假设下的概率：

w_{XY}	0	1	2	3	4
w_Y	3	4	5	6	7
$P(W_{XY}=w_{XY})$ $=P(W_Y=w_Y)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

从上表可以看见，W_Y 的分布关于$\frac{n(N+1)}{2}$ 对称，这一点也可从它的定义得出，即下面的定理.

定理 4.2 Wilcoxon 秩和统计量的零分布关于 $\mu=\frac{n(N+1)}{2}$ 对称.

证明 设 $R^*=(Q_1,\cdots,Q_m,R_1,\cdots,R_n)$ 为全样本的秩统计量. 由定理 3.3 知，R^* 在 $\mathscr{R}=\{(i_1,\cdots,i_N)$ 是$(1,\cdots,N)$ 的排列$\}$ 上均匀分布. 又由于在 H_0 下，$-X_1,\cdots,-X_m,-Y_1,\cdots,-Y_n$ 也是 iid 的，而它的秩统计量为$(N-Q_1,\cdots,N-Q_m,N-R_1,\cdots,N-R_n)$，故在 H_0 下

$$(Q_1,\cdots,Q_m,R_1,\cdots,R_n)\overset{d}{=}(N-Q_1,\cdots,N-Q_m,N-R_1,\cdots,N-R_n)$$

由此可知，在 H_0 下，$\sum_{i=1}^{n}R_i\overset{d}{=}n(N+1)-\sum_{i=1}^{n}R_i$. 因此在 H_0 下

$$W-\frac{n(N+1)}{2}\overset{d}{=}\frac{n(N+1)}{2}-W$$

于是由定理 3.1 知，W 的零分布关于其期望 $\mu=\frac{n(N+1)}{2}$ 对称. □

利用这一对称性，对其造表是很有用的，当然，此定理也适用于检验统计量 W_{XY}. 但是以上的简单方法对于大的样本就显得麻烦了. 下面引进的方法则可以写成简单的计算机程序来计算 W_{XY} 的分布. 关键是计算上表中 W_{XY} 等于某一值的数目. 定义

$$\overline{P}_{m,n}(k)=\#(W_{XY}=k)$$

比如，在上表中 $\overline{P}_{2,2}(2)=\#(W_{XY}=2)=2$ 及 $\overline{P}_{2,2}(0)=\#(W_{XY}=0)=1$. 下面用两个定理的形式描述我们的方法.

定理 4.3 如 X 和 Y 样本的大小分别是 m 和 n，

$$\overline{P}_{m,n}(k)=\overline{P}_{m,n-1}(k-m)+\overline{P}_{m-1,n}(k)$$

这里，对 $k<0$，$\overline{P}_{i,j}(k)=0$；对 $k=0$，$\overline{P}_{i,0}(k)=\overline{P}_{0,j}(k)=0$；对 $k\neq 0$，$\overline{P}_{i,0}(k)=\overline{P}_{0,j}(k)=1$.

证明 对于 $W_{XY}=k$ 有两种情况：1. 上表相应的列以 X 结尾. 2. 以 Y 结尾. 对于情况 1，按

W_{XY} 定义，去掉这个 X，不改变 W_{XY} 的值（还是 k），仅把 X 的样本大小由 m 变为 $m-1$，这就产生了上式的第二项. 对于情况 2，按 W_{XY} 定义，去掉这个 Y，W_{XY} 的值会减少 m 而变成 $k-m$（因为这个 Y 比 m 个 X 大，所以对 W_{XY} 的贡献是 m），而 Y 的样本大小由 n 变为 $n-1$，这就产生了上式的第一项. □

定理 4.4 在 $H_0:\theta=0$ 下，记概率 $P_{m,n}(k)=P_{H_0}(W_{XY}=k)$，则有

$$P_{m,n}(k)=\frac{n}{m+n}P_{m,n-1}(k-m)+\frac{m}{m+n}P_{m-1,n}(k)$$

这里，对 $k<0, P_{i,j}(k)=0$；对 $k=0, P_{i,0}(k)=P_{0,j}(k)=0$；对 $k\neq 0, P_{i,0}(k)=P_{0,j}(k)=1$.

证明 在 $H_0:\theta=0$ 下，一共有 $\binom{m+n}{m}$ 个等可能的 X 和 Y 组成的序列，因此

$$P_{m,n}(k)=\frac{\overline{P}_{m,n}(k)}{\binom{m+n}{m}}=\frac{m!n!}{(m+n)!}[\overline{P}_{m,n-1}(k-m)+\overline{P}_{m-1,n}(k)]$$

$$=\frac{n}{m+n}\frac{\overline{P}_{m,n-1}(k-m)}{\binom{m+n-1}{m}}+\frac{m}{m+n}\frac{\overline{P}_{m-1,n}(k)}{\binom{m+n-1}{n}}$$

$$=\frac{n}{m+n}P_{m,n-1}(k-m)+\frac{m}{m+n}P_{m-1,n}(k)$$ □

由此定理，我们可用递推的方法计算 $P(W_{XY}\leqslant a)$. 且给出了当 $k<0, k=0$ 及 $i=0, j=0$ 时 $P_{i,j}(k)$ 的初始值，由此及上面递推公式可依次计算 $P_{1,1}(0), P_{1,1}(1), \cdots, P_{2,1}(0), P_{2,1}(1), \cdots$ 的值，则 $P(W_{XY}\leqslant a)$ 的零分布表可由此算出（见附表 5）. 注意：当有 X 及 Y 的值相同时，只需定义

$$W_{XY}=\#(X_i<Y_j, i\in I_m, j\in I_n)+\#(X_i=Y_j, i\in I_m, j\in I_n)$$

即可.

例 4.2 一名熟练工人先后用两台机床加工同样的产品. 现从这两台机床加工的产品中随机地抽取若干产品，测得产品直径为（单位：mm）

甲	18.1	17.7	17.2	19.1	17.0	17.5	17.8	18.7
乙	18.3	19.0	18.9	17.3	16.9	18.4	17.6	18.6

试问甲、乙两台机床加工的产品的平均直径有无显著差异？

对于本例的数据，其 Y 样本的秩和为 $W_Y=73$，如显著性水平 $\alpha=0.1$，则由附表 5 查得 $c_1=16, c_2=mn-c_1=48$，由此可知，没有理由拒绝 H_0.

在大样本时，容易证明

定理 4.5 若 $m,n\to\infty$，同时有 $\frac{m}{(m+n)}\to\lambda, 0<\lambda<1$，则在 $H_0:\theta=0$ 下，渐近地有

$$\frac{W_{XY}-\frac{mn}{2}}{\sqrt{\frac{mn(N+1)}{12}}}\xrightarrow{\mathscr{L}}N(0,1),\qquad \frac{W_Y-\frac{n(N+1)}{2}}{\sqrt{\frac{mn(N-1)}{12}}}\xrightarrow{\mathscr{L}}N(0,1)$$

证明 可见本章阅读知识一节，也可见第八章.

对于上面的正态近似，我们可以写成如下形式：

$$P(W_Y \leqslant k) \approx \Phi\left(\frac{k - \dfrac{n(N+1)}{2}}{\sqrt{\dfrac{mn(N+1)}{12}}}\right)$$

$$P(W_Y \leqslant k) \approx \Phi\left(\frac{k + 0.5 - \dfrac{n(N+1)}{2}}{\sqrt{\dfrac{mn(N+1)}{12}}}\right)$$

其中最后一个近似用了连续修改.对于上面两个近似结果的好坏，Lehmann([14])给出了下面的几组数据：

$m = 3, n = 16$

k	6	7	8	9	10
精确值	0.012	0.024	0.048	0.083	0.131
无连续修改	0.010	0.019	0.035	0.061	0.098
有连续修改	0.014	0.026	0.047	0.078	0.123

$m = 4, n = 12$

k	13	15	20	23	25
精确值	0.004	0.010	0.025	0.106	0.158
无连续修改	0.005	0.011	0.045	0.091	0.138
有连续修改	0.006	0.012	0.051	0.102	0.151

$m = 8, n = 8$

k	44	46	48	52	56	58
精确值	0.005	0.010	0.019	0.052	0.117	0.164
无连续修改	0.006	0.010	0.018	0.047	0.104	0.147
有连续修改	0.007	0.012	0.020	0.052	0.114	0.159

由上面的数字计算可以看出，即使 m, n 较小，使用带连续性修改后的正态近似计算还是可行的.一般来说，使用连续修改后的近似要比未作修改的近似来得好，但是当 W_Y 取两端值时，情况却未必如此.

注：在 X 及 Y 有的值相同时，即全样本有结时，如用 $(\tau_1, \cdots, \tau_g)$ 表示全样本的结统计量，则可以证明 W_Y, W_{XY} 零分布的期望和方差为：

$$E_{H_0}(W_Y) = \frac{n(N+1)}{2}, \qquad E_{H_0}(W_{XY}) = \frac{mn}{2}$$

$$\mathrm{Var}_{H_0}(W_Y) = \mathrm{Var}_{H_0}(W_{XY}) = \frac{mn(N+1)}{12} - \frac{mn\sum_{i=1}^{g}(\tau_i^3 - \tau_i)}{12N(N-1)}$$

其证明可见[14]. 同样可以证明，当 $\min(m,n) \to \infty$ 时，有

$$\frac{W_Y - E_{H_0}(W_Y)}{\sqrt{\mathrm{Var}_{H_0}(W_Y)}} \xrightarrow{\mathscr{L}} N(0,1)$$

$$\frac{W_{XY} - E_{H_0}(W_{XY})}{\sqrt{\mathrm{Var}_{H_0}(W_{XY})}} \xrightarrow{\mathscr{L}} N(0,1)$$

对于上面的正态近似，不如无结时的近似好，但是 Lehmann 于 1961 年指出，上面的近似还是可行的(见 JASA. 1961, 56: 293—298).

例 4.3 在心理咨询的研究中，假设随机地抽取 80 人，从中随机地抽取 40 人给予心理咨询，而剩下的 40 人没有心理咨询. 之后，对每个人的心理状态进行测试，测试结果分为好、尚好、较差和差四种，其人数为

咨询与否	差	较差	尚好	好	总和
Y：有	5	7	16	12	40
X：无	7	9	15	9	40

对这组数据，如我们假设某总体只取 4 个值：1(代表差)、2(代表较差)、3(代表尚好)、4(代表好)，则一共有 $5 + 7 = 12$ 个取最小的；$7 + 9 = 16$ 个取次最小的；$16 + 15 = 31$ 个取次最大的；$12 + 9 = 21$ 个取最大的，这四个值的秩分别为

$$\frac{1 + 2 + \cdots + 12}{12} = 6.5, \quad \frac{13 + 14 + \cdots + 28}{16} = 20.5$$

$$\frac{29 + 28 + \cdots + 59}{31} = 44, \quad \frac{60 + 61 + \cdots + 80}{21} = 70$$

则 Wilcoxon 秩和统计量

$$W = 5 \times 6.5 + 7 \times 20.5 + 16 \times 44 + 12 \times 70 = 1720$$

又因为

$$m = n = 40, \quad E(W) = 1620, \quad \tau_1 = 12, \quad \tau_2 = 16$$

$$\tau_3 = 31, \quad \tau_4 = 21, \quad \sqrt{\mathrm{Var}(W)} = 99.27$$

所以检验的 p 值为 $P(W \geqslant 1720) = 1 - \Phi(1.01) = 0.16$，故对于 $\alpha = 0.1$ 来说，我们没有理由拒绝 H_0.

从定义可以看出，W_{XY} 实际上是基于 mn 个差 $Y_i - X_j$ 的符号统计量，我们可用 Hodges-Lehmann 估计量

$$\hat{\theta} = \underset{i,j}{\mathrm{median}}(Y_i - X_j)$$

来估计 θ. 令 $D_{(1)}, \cdots \leqslant D_{(mn)}$ 表示按升幂排列的 $(Y_i - X_j), i \in I_m, j \in I_n$ 的值. 如果 k 满足 $P_{H_0}(W_{XY} \leqslant k) = \frac{\alpha}{2}$，则 θ 的 $(1 - \alpha)100\%$ 置信区间为

$$[D_{(k+1)}, D_{(mn-k)})$$

对大样本情况，近似地有

$$k = \frac{mn}{2} - 0.5 - Z_{\frac{\alpha}{2}} \sqrt{\frac{mn(m + n + 1)}{12} - \frac{mn \sum_{i=1}^{g} (\tau_i^3 - \tau_i)}{12(m + n - 1)(m + n)}}$$

相应于更广泛的检验统计量，类似 §3.4 中所讲的，我们也可以定义一般的 Hodges-Lehmann 估计如下：假设 $V(X_1,\cdots,X_m;Y_1,\cdots,Y_n)$ 是一个关于零假设 $H_0:\theta=0$ 的检验统计量，其零分布关于某 ξ 对称，且对于任意的 $(X_1,\cdots,X_m;Y_1,\cdots,Y_n)$，$V(X_1,\cdots,X_m;Y_1+h,\cdots,Y_n+h)$ 关于 h 是非增的. 定义

$$\theta^* = \sup\{\theta:V(X_1,\cdots,X_m;Y_1-\theta,\cdots,Y_n-\theta)>\xi\}$$
$$\theta^{**} = \inf\{\theta:V(X_1,\cdots,X_m;Y_1-\theta,\cdots,Y_n-\theta)<\xi\}$$

则 θ 的基于检验统计量 V 的 Hodges-Lehmann 估计为

$$\hat{\theta}=\frac{\theta^*+\theta^{**}}{2}$$

如果取

$$V=\frac{\sqrt{\frac{mn}{m+n}}(\bar{Y}_n-\bar{X}_m)}{S_{mn}}$$

则 θ 的 HL 估计为 $\bar{Y}_n-\bar{X}_m$；如果取 V 为 Wilcoxon 秩和统计量，则此时 θ 的 HL 估计即为前面给出的 $\text{median}_{i,j}(Y_i-X_j)$

关于 Hodges-Lehmann 估计的确切分布，请参看习题 2 和习题 3. 关于其分布的细节，请参看[7] §7.2；关于其极限分布，请参见习题 4，细节请参看[7] §7.3.

虽说定理 4.1 给出了 R_i 的零分布，但是在考虑检验的功效时，却需要知道其在备选假设下的分布，此时我们有下面的结论. 令随机样本 $X_1,\cdots,X_m$ 及 $Y_1,\cdots,Y_n$ 分别来自于分布为 $G(x)$ 和 $H(y)$ 的样本，这里 $G(x)$ 和 $H(y)$ 是任意的绝对连续分布函数，并以 $g(x)$ 和 $h(y)$ 记其密度. 记 $R_{(1)}<\cdots<R_{(n)}$ 为顺序统计量 $Y_{(1)}<\cdots<Y_{(n)}$ 在 X 和 Y 的混合样本中的(次序)秩. 我们有如下定理(不证明)：

定理 4.6　假设由 $h(x)>0$ 可导出 $g(x)>0$，则

$$P(R_{(1)}=r_1,\cdots,R_{(n)}=r_n)=\frac{1}{\binom{m+n}{m}}E\left[\prod_{i=1}^{n}\frac{h(V_{(r_i)})}{g(V_{(r_i)})}\right]$$

这里 $V_{(r_1)}<\cdots<V_{(r_n)}$ 为来自分布 G 的一个大小为 $m+n$ 的样本之顺序统计量 $V_{(1)}<\cdots<V_{(m+n)}$ 中的一部分.

证明见[14].

特别地，对于 $F\in\Omega_c$(密度为 f)，我们考虑 $G(x)=F(x)$ 及 $H(y)=F(y-\theta)$ 的情况. 显然，在零假设 $H_0:\theta=0$ 下，

$$P(R_{(1)}=r_1,\cdots,R_{(n)}=r_n)=\frac{1}{\binom{m+n}{m}}$$

也就是说 $R_{(1)}<\cdots<R_{(n)}$ 是均匀分布的，即对于 $\binom{m+n}{m}$ 种组合是等概率的.

我们在第八章将看到定理 4.6 的应用.

上面我们考虑了位置参数两样本问题的几个检验统计量，这几种方法之间的优劣又如何

分布F中抽出的次序样本，且与$X_{(1)},\cdots,X_{(n_1)}$独立。于是，在表达式

$$P(Y_{(s)}-X_{(r)}\geqslant\theta)=P(Z_{(s)}\geqslant X_{(r)})$$

中，通过固定$X_{(r)}=x$求$Z_{(s)}\geqslant X_{(r)}$的条件概率，用(2·3)式，得

$$\begin{aligned}P(Z_{(s)}\geqslant X_{(r)}|X_{(r)}=x)&=P(Z_{(s)}\geqslant x)\\&=\sum_{i=0}^{s-1}\binom{n_2}{i}F^i(x)(1-F(x))^{n_2-i},\end{aligned}$$

再注意到$X_{(r)}$有分布函数（2·5），而$dF_r(x)=\dfrac{n_1!}{(r-1)!(n_1-r)!}$
$\cdot F^{r-1}(x)(1-F(x))^{n_1-r}dF(x)$，得

$$\begin{aligned}P(Y_{(s)}-X_{(r)}\geqslant\theta)&=P(Z_{(s)}-X_{(r)}\geqslant 0)\\&=\int_{-\infty}^{\infty}\sum_{i=0}^{s-1}\binom{n_2}{i}F^i(x)(1-F(x))^{n_2-i}\frac{n_1!}{(r-1)!(n_1-r)!}\\&\qquad\cdot F^{r-1}(x)(1-F(x))^{n_1-r}dF(x)\\&=r\binom{n_1}{r}\sum_{i=0}^{s-1}\int_0^1 t^{r+i-1}(1-t)^{n-r-i}dt\binom{n_2}{i}\quad(n=n_1+n_2)\\&=r\binom{n_1}{r}\sum_{i=1}^{s-1}\binom{n_2}{i}\frac{(r+i-1)!(n-r-i)!}{n!}\\&=\sum_{i=0}^{s-1}\binom{r+i-1}{r-1}\binom{n-r-i}{n_1-r}\Big/\binom{n}{n_1}\\&=\frac{n_1}{n}\sum_{i=0}^{s-1}\binom{r+i-1}{r-1}\binom{n-r-i}{n_1-r}\Big/\binom{n-1}{n_1-1}.\end{aligned}\qquad(2\cdot 93)$$

此式和号下各项为超几何概率。在n不很大时，可由超几何分布表查得。令此式等于$1-\alpha$以决定r，s。则$[X_{(r)},Y_{(s)}]$就是θ的置信系数为$1-\alpha$的置信区间。当然，由一个等式不能决定两个未知量r，s，这可以采用下述由直观提供的想法：令

$$r=[(n_1+1)/2]-l,\quad s=[(n_2+1)/2]+l\quad(2\cdot 94)$$

这表示分别从足标$[(n_1+1)/2]$和$[(n_2+1)/2]$出发，足标上移的距离相同。以(2·94)中的r，s代入（2·93）用“try and error”的方法决定l，使（2·93）正好等于$1-\alpha$。如果不存在这样的l，

则或者修改α，或者对相邻的两个l使用随机化手续，如近在§2.4的三段中所做的那样。

当n较大时，使用上述精密的方法去决定l可能变得过于繁复而不可行。这时可以用正态分布逼近超几何分布的方法，以决定l的近似值。这在概念上与依据正态逼近（2·88）来决定r的近似值（2·89）相同，但在推导上要麻烦得多。此处我们不给出有关细节，而只将结果写出：

$$r \approx [(n_1+1)/2] - \frac{\sqrt{n_1 n}}{2(\sqrt{n_1}+\sqrt{n_2})}u_\alpha, \qquad (2\cdot 94)$$

$$s \approx [(n_2+1)/2] + \frac{\sqrt{n_2 n}}{2(\sqrt{n_1}+\sqrt{n_2})}u_\alpha. \qquad (2\cdot 95)$$

用类似的方法，可求得θ的形如$Y_{(s)}-X_{(r)}$的置信系数$1-\alpha$的置信下限。所不同的是：在近似公式（2·94）中，右边的减号要改为加号，而在（2·95）中，两项相加改为相减。

如要求θ的置信系数$1-\alpha$的置信区间，则一个近似的作法是以前所指出过的：先用上述方法找到θ的置信系数$1-\alpha/2$的置信上、下限$Y_{(s)}-X_{(r)}$及$Y_{(s')}-X_{(r')}$，于是$[Y_{(s')}-X_{(r')}, Y_{(s)}-X_{(r)}]$作为$\theta$的置信区间，其置信系数至少为$1-\alpha$。若要得到确切置信系数的解，也可从头开始，用次序统计量的分布理论，依照得出（2·93）类似的推理去做。此法在原则上虽可行，但过于繁复因而缺乏实用价值。故我们也不去讲究其细节了。

最后考虑有关θ的假设检验问题。所指的是形如

$$H_1:\theta=\theta_0,\quad H_2:\theta\leqslant\theta_0,\quad H_3:\theta\geqslant\theta_0. \qquad (2\cdot 96)$$

这样的原假设，各有相应的对立假设。例如，H_1的对立假设为$K_1:\theta\neq\theta_0$。

这种问题可以用以下一些方法去解决。

1. 用大样本理论，从θ的具有渐近正态性的点估计出发。由于渐近方差将依赖于总体分布F，它需要据样本去估计之。于是渐近分布取代精确分布所产生之误差以外，又加上了用样本方差

估计值取代样本方差所产生之误差。故除非样本大小n很大，这种方法的效果不甚理想。一般只是在不得已时偶一用之。

2. 根据置信区间与假设检验的一般关系，通过用前述方法作θ的形如$[Y_{(s')}-X_{(r')},Y_{(s)}-X_{(r)}]$的置信区间(或置信上、下限)，用之检验原假设($H_1$或$H_2$，$H_3$)。其具体作法与§2.4(三)段结尾处所描述的类似。

3. 现在介绍一种也是基于次序统计量的作法。不失一般性，可设(2·96)中的$\theta_0=0$。因为只须用$Y_i'=Y_i-\theta_0$代替$Y_i(i=1,\cdots,n_2)$即可做到这一点。如前以$Z_{(1)}\leqslant\cdots\leqslant Z_{(n)}(n=n_1+n_2)$，记合样本$(X_1,\cdots,X_{n_1},Y_1,\cdots,Y_{n_2})$的次序统计量，而$m$记其样本中位数。为确定计暂设$n$为偶数且$m$不等于$Z_{(1)},\cdots,Z_{(n)}$中任一个。于是$Z_{(1)},\cdots,Z_{(n/2)}$在$m$的左边，其余的在右边。当原假设$\theta=0$成立时，全体$n$个样本$X_1,\cdots,Y_{n_2}$是独立同分布，其地位完全平等。故$X_1,\cdots,X_{n_1}$这$n_1$个点中，近似地应有$n_1/2$个点在$m$的左边。以$\eta$记$X_1,\cdots,X_{n_1}$中落在$m$左边的个数。若$|\eta-n_1/2|$过大，则这将是$\theta=0$不成立的一种启示。

显然，当原假设$\theta=0$成立时，η服从超几何分布：

$$P(\eta=i)=\binom{n_1}{i}\binom{n_2}{n/2-i}/\binom{n}{n/2}.$$

因此，在原假设$\theta=0$成立之下，若$\eta=a<n_1/2$，则得到像a这样的偏离中心$n/2$，之值，或比之偏离更大之值，其概率应为

$$p_a=\sum_{i\leqslant a}P(\eta=i)+\sum_{i\geqslant n_1-a}P(\eta=i)$$

$$=2\sum_{i\leqslant a}\binom{n_1}{i}\binom{n_2}{n/2-i}/\binom{n}{n/2}. \qquad (2\cdot97)$$

如果p_a小于给定的检验水平α，则否定原假设$\theta=0$。当$a\geqslant n_1/2$时，(2·97)式中的a改为$n-a$，检验规则不变。

如果要检验的原假设是$\theta\leqslant0$，则显然只有η的大值(远大于$n_1/2$的那些值)才是背离原假设的。考虑到在一般应用中，给定的检验水平α都远低于0.5。故若$\eta=a$，而$a\leqslant n_1/2$，则必然接受原

假设。若$a>n_1/2$，则计算

$$\widetilde{p}_a=\sum_{i\leqslant n_1-a}\binom{n_1}{i}\binom{n_2}{n/2-i}\Big/\binom{n}{n/2}\left(=p_a/2\right). \qquad (2\cdot 98)$$

当$\widetilde{p}<\alpha$时否定原假设$\theta\leqslant 0$.检验原假设$\theta\geqslant 0$的方法与此类似，建议读者仔细写出来。

如果$n=n_1+n_2$为奇数，则合样本的样本中位数m是这n个样本中之一。把这一个挑去不算，而按n为偶数的情况处理，也就是说，若m为X样本之一，则n_1改为n_1-1而n_2不动；若m为Y样本之一，则n_1不动而n_2改为n_2-1.n当然改为$n-1$.其余一切按上述处理。

当n较小时，p_a或q_a之值可以通过查超几何分布表得到.这里有一个与查表方便有关之点值得指出.记$a_0=\min(a,n_1-a)$。则在计算p_a或q_a时，涉及的和为$\sum_{i\leqslant a_0}$。这是从考察X样本的角度来看的。如若考察Y样本中落在m的左侧的个数b，则也可作出检验（这检验当然与已给出的检验等价）.记$b_0=\min(b,n_2-b)$，这检验需计算的概率为

$$\widetilde{p}_{b_0}=2\sum_{i\leqslant b_0}\binom{n_1}{n/2-i}\binom{n_2}{i}\Big/\binom{n}{n/2}, \qquad (2\cdot 99)$$

或

$$\widetilde{q}_{b_0}=\sum_{i\leqslant b_0}\binom{n_1}{n/2-i}\binom{n_2}{i}\Big/\binom{n}{n/2}. \qquad (2\cdot 100)$$

如果$a_0<b_0$，则以用（2·97）或（2·98）为方便；若$a_0>b_0$，则以用（2·99）或（2·100）为方便。

当n较大时，查表或直接计算往往不可行.这时可考虑用正态分布逼近超几何分布。例如，

$$p_{b_0}=\sum_{i\leqslant a}\binom{n_1}{i}\binom{n_2}{n/2-i}\Big/\binom{n}{n/2}\approx\Phi(t_0),$$

其中

$$t_0=\sqrt{2n}\left(a_0+\frac{1}{2}-\frac{n_1}{2}\right)\Big/\sqrt{n_1n_2}\ .$$

分子中的1/2是考虑到连续性修正。

§2·6 连续分布的容忍限与容忍区间

设$X_1,\cdots,X_n$是随机变量X的简单样本，X的分布F处处连续。为直观计，不妨把X看成某批量生产的产品之一项质量指标，且假定X之值愈低，则该产品之质量愈不好。我们希望回答这样的问题：绝大部分产品的质量指标的低限（下限）如何。更确定地说，指定一个很小的概率β，如$\beta=0.05$。我们要从样本$X_1,\cdots,X_n$算出一个低限$I=I(X_1,\cdots,X_n)$，使产品质量指标不超过I的那部分最多只占$100\beta\%$，即

$$F(I(X_1,\cdots,X_n))\leqslant\beta.$$

但因样本有随机性，一般说来，不论你如何去选择统计量I，也不可能万无一失地保证（2.101）必成立。因此，我们只能以一定的概率$1-\gamma$（通常$\gamma>0$很小）保证（2·101）成立，即要求

$$P(F(I(X_1,\cdots,X_n))\leqslant\beta)\geqslant 1-\gamma. \qquad (2\cdot102)$$

如果对给定的（β，γ），某统计量I满足（2·102），则称它是总体分布F的（β,γ）容忍下限。"容忍"（Tolerance）一词用于此处其义似觉费解，后面将试图作一点说明。

在已知总体分布为正态时，根据显然的理由，人们去寻求形如$\overline{X}-cS$的容忍下限，此处$\overline{X}$和S^2分别是$X_1,\cdots,X_n$的样本均值和样本方差。于是问题归结为：根据给定的（β,γ）和n去决定c。这问题属于参数统计范围，在一般教程中多有论述，c之值也有表可查。

此处我们对分布F除连续外别无其他限制，问题属于非参数范围。以$X_{(2)}\leqslant\cdots\leqslant X_{(n)}$记次序统计量。我们设想：开头几个即$X_{(1)}$，$X_{(2)}$，…应接近于质量指标的低限，因此考虑形如$X_{(r)}$的容忍下限。据定理2·1，然后将（2·3）用于$F\sim R(0,1)$的情况，得

$$P(F(X_{(r)}))\leqslant\beta=\sum_{i=r}^{n}\binom{n}{i}\beta^i(1-\beta)^{n-i}.$$

此式与（2.102)结合，得出确定 r 的关系式：

$$\sum_{i=0}^{r-1}\binom{n}{i}\beta^i(1-\beta)^{n-i}\leqslant\gamma. \qquad (2\cdot103)$$

可以使用二项分布表，用“try and error”的方法去决定一个满足此式的最大的 r。然后，取 $X_{(r)}$ 作为容忍下限。常见的情况是存 r_0，使

$$\sum_{i=0}^{r_0-1}\binom{n}{i}\beta^i(1-\beta)^{n-i}<\gamma<\sum_{i=0}^{r_0}\binom{n}{i}\beta^i(1-\beta)^{n-i}.$$

这时，容忍下限 $X_{(r_0-1)}$ 的保证概率将略大于 $1-\gamma$。

只要 n 较大，则二项概率一致地很小，因而这个差在应用上也许不太重要。当 n 很大时，可以通过正态逼近去决定 r。

类似地可给出 (β,γ)--容忍上限 $J=J(X_1,\cdots,X_n)$ 的定义

$$P(F(J(X_1,\cdots,X_n))\geqslant1-\beta)\geqslant1-\gamma.$$

若寻求 $X_{(s)}$ 型的容忍上限，则与下限类似，导出 s 满足的关系为

$$\sum_{i=0}^{n-s}\binom{n}{i}\beta^i(1-\beta)^{n-i}\leqslant\gamma.$$

使用二项分布表，用“try and error”的方法，去决定一个满足此式的最小的 s。然后取 $X_{(s)}$ 作为容忍上限。与下限相似，一般保证概率略大于 $1-\gamma$。

以上考虑的是容忍限，现考虑容忍区间。所谓 $[I,J]$ 是总体分布 F 的 (β,γ) 容忍区间，是指

$$P(F(J)-F(I)\geqslant1-\beta)\geqslant1-\gamma. \qquad (2\cdot104)$$

找容忍区间的一个办法是通过找容忍限。容易证明：若 J 和 I 分别是 F 的 $(\beta/2,\ \gamma/2)$ 容忍上、下限，则 $[I,J]$ 就是 F 的 (β,γ) 容忍区间。因此，可按前述方法构造出形如 $[X_{(r)},X_{(s)}]$ 的容忍区间。但这样定出的容忍区间一般偏于“保守”，特别是 n 并不很大时，就是说，r 比实际所需的小而 s 比实际所需的大。但使用次序统计量的理论，也不难求出形如 $[X_{(r)},X_{(s)}]$ 的精确解。

以$U_1\leqslant\cdots\leqslant U_n$记$R(0,1)$中抽出的大小为$n$的次序样本，则不难证明：若$1\leqslant r<s\leqslant n$，则$U_s-U_r$与$U_{s-r}$同分布。我们把这个简单事实的证明留给读者。这样，利用定理2·1，并记$k=s-r$，得

$$P(F(X_{(s)})-F(X_{(r)})\geqslant 1-\beta)=P(U_s-U_r\geqslant 1-\beta)$$
$$=P(U_k\geqslant 1-\beta).$$

让此概率等于$1-\gamma$，得到决定k的关系式

$$\sum_{i=0}^{n-k}\binom{n}{i}\beta^i(1-\beta)^{n-i}\leqslant\gamma, \qquad (2\cdot105)$$

由此式决定了k以后，再根据$s-r=k$决定s和r。这就必须引入另一条件。依对称性考虑，可把此条件定为$s+r=n$。除非n，k同奇偶，此两式定不出整数s,r。在这种情况下，可以把第二个条件改为$s+r=n+1$或$n-1$。

通常，对一种产品指标定下了一个规格区间$[a,b]$，只有当产品的质量指标X落在$[a,b]$内时，这产品才是合格的。可以把$[a,b]$这个区间看成是质量指标波动所能容忍的限度。现在要问这样一个问题：全部产品中有多大的部分（百分率）其指标在a b之间？从直接的意义看，这本是一个估计问题，即估计一个与总体分布有关的量$\theta=F(b)-F(a)$。不难作出θ的点估计及大样本区间估计，但具有确切置信系数的小样本区间估计则不易求得。受前面讨论的启发，我们可以换一个角度来看这个问题。设有了次序样本$X_{(1)}\leqslant\cdots\leqslant X_{(n)}$，并为说明简便计，设存在$r,s$，使恰有$X_{(r)}=a,X_{(s)}=b$，然后用$F(x_{(s)})-F(X_{(r)})$取代$F(b)-F(a)$。给定$\gamma$，$0<\gamma<1$。要找出$\beta$，使

$$P(F(X_{(s)})-F(X_{(r)})\geqslant 1-\beta)\geqslant 1-\gamma. \qquad (2\cdot106)$$

β找到后，再把$F(X_{(s)})-F(X_{(r)})$还原成原来的$F(b)-F(a)$。这样，（2·106）式就可以解释为："可以用$1-\gamma$的概率保证符合规格的产品比率至少为$100(1-\beta)\%$"。这个解释对应用者来说大概会觉得可以理解。对注意理论的人来说，他当然会看到上述

推理中的不合逻辑之处，问题在于：$F(b)-F(a)$虽然未知，但它是一确定的常数，并无随机性。它或者$\geqslant 1-\beta$，或者$<1-\beta$。也就是说，它$\geqslant 1-\beta$的概率只能为1或0，说它是$1-\gamma$是没有意义的。

上面这段讨论也就多少说明了“容忍区间”一词中，“容忍”的意义何在，虽然这个名词的确切性仍可以讨论。

§2·7 极值方法

所谓极值方法，是指那些统计方法，其使用只涉及样本中的最大值或(和)最小值。有的极值方法也用到若干个次大值或次小值。

从模型的原本上说，极值方法可认为是非参数性的。但在应用上，往往先依据定理2·3这一类的极值分布定理，把模型过渡成参数型的(即极值分布三种类型之一)。从这个角度看，把极值方法归入参数统计也言之有理。我们并不需要在此对这个问题下一决断。只是由于这个内容在其他统计课程中也没有适当安排，而它又确是次序统计量的一种应用，故在这一章中作点简略的介绍，其详可参看有关著作，如Gumbel的专著《Statistics of Extremes》(Columbia University Press, 1958)。

极值统计应用于这样的情况：在其中我们最关心的是变量观察值的极端值。例如在一个地震多发区，逐日地震发生频繁，但绝大多数震级都很低，无关紧要，关心的是在一定时期(一天、一个月等)中地震最大震级。一条河流在某处的水位逐日有变，关心的是其在一定时期的最高水位或最低水位，前者与防汛有关，而后者与航运有关。这类例子可举出很多。

问题的一般模式如下：有一个我们关心的随机变量X，其分布F未知。对其进行了若干次观察，得到样本$X_1, \cdots, X_N$。按某种方式把它们分成大小为n的组：

第一组：$X_1, \cdots, X_n$,

第二组：$X_{n+1}, \cdots, X_{2n}$,

……………

第m组　$X_{(m-1)n+1}, \cdots, X_{mn}$.

分组的原则依数据的性质及其他考虑而定。如在地震数据中，每一组可以是一年365天逐日最大震级的记录。在材料试验中，按一定的方法把试验样品分组，每组内各样品的断裂强度数据即构成上面的一组数据。按这样分组时，可能$N>mn$，即有些观察值可能用不上，那也没有办法。但如试验是由人安排的，这种情况总可设法避免。由于要使用极值定理，n不可太小。又为了估计极值分布中的未知参数，m也不可太小。这就要求资料数量有一定的规模。

以Y_i记上述第i组数据中的最大值(为确定计，此处我们考虑最大值，最小值的处理类似)，得$Y_1, \cdots, Y_m$。假定全部mn个原始数据$X_1, \cdots, X_{mn}$是独立同分布的，则$Y_1, \cdots, Y_m$也是独立同分布。又假定n已足够大，且总体分布F适合定理2·3的条件，则存在常数u和$\alpha>0$，使对每个i，$\alpha(Y_i-u)$的分布函数近似地为$\exp(-e^{-x})$，这样，Y_i的分布可认为是$\exp(-e^{-\alpha(x-u)})$。

如果知道了u和α，就可以回答一些感兴趣的问题。例如，算出x_0使

$$\exp(-e^{-\alpha(x_0-u)})=0.99. \qquad (2\cdot107)$$

则事件$\{Y_i<x_0\}$的概率只有百分之一。如果Y_i是一年内所记录的最大震级，则x_0可解释为：在指定的一年中碰到震级超过x_0的地震，其机会不过百分之一。通常把这个x_0说成是“百年一遇”的地震震级。这个数据在建设一项大型工程时有参考意义。但通常参数u和α都未知，需要通过样本进行估计。这就是为什么我们必须有m组数据(m不太小)，以得到$Y_1, \cdots, Y_m$，它们看成是从具有分布$\exp(-e^{-\alpha(x-u)})$的总体中抽出的简单样本，据此估计u和α。方法很多，较重要的有以下几种：

1．样本分位数法

把（2·107）左端的函数记为$G(x)$，得$G(u)=e^{-1}=0.3679=p_1$.换句话说，u是分布G的p_1分位数.因为$Y_1,\cdots,Y_m$是G的简单样本，故u可通过$Y_1,\cdots,Y_m$的样本p_1分位数去估计之.其次，有

$$G(u+\frac{1}{a})=\exp(-e^{-1})=0.6922=p_2,$$

故$u+1/\alpha$可通过$Y_1,\cdots,Y_m$的样本p_2分位数去估计之．这与u的估计结合，即得出α的估计．

这个方法简单易行.在m很大时效果也好.但m较小时，由于样本分位数的多值性难以妥善处理，效果就会差些．故这时不宜采用此法．

2．最小二乘法(线性回归法)

以$Y_{(1)}\leqslant\cdots\leqslant Y_{(m)}$记$Y_1,\cdots,Y_m$的次序统计量.按定理2·1，$G(Y_{(1)}),\cdots,G(Y_{(m)})$是均匀分布$R(0,1)$的次序样本.易算出

$$E(G(Y_{(i)}))=i/(m+1),\quad i=1,\cdots,m,$$

故在m较大时，可以把$i/(m+1)$作为$G(Y_{(i)})$的近似值.这可以由$G(Y_{(i)})$的方差

$$\text{Var}(G(Y_{(i)}))=\frac{i(m+1-i)}{(m+1)^2(m+2)}$$

当m较大时甚小看出.于是可写

$$\exp(-\exp(-\alpha(Y_{(i)}-u)))\approx\frac{i}{m+1},\quad i=1,\cdots,m,$$

取两次对数，得

$$\alpha(Y_{(i)}-u)\approx-\log(-\log\frac{i}{m+1})=c_i,\ i=1,\cdots,m.\quad(2\cdot108)$$

此式只是近似的．如强令其相等，则当$m>2$时，得出矛盾方程组．只好用最小二乘法来处理，即求u与α之值$\hat{u}$，$\hat{a}$，使表达式

$$\sum_{i=1}^{m}(\alpha(Y_{(i)}-u)-c_i)^2$$

达到最小，不难解出结果为

$$\hat{\alpha}=\sum_{i=1}^{m}c_i(Y_{(i)}-\bar{Y})/\sum_{i=1}^{m}(Y_{(i)}-\bar{Y})^2 \quad (\bar{Y}=\sum_{i=1}^{m}Y_i/m),$$
$$\hat{u}=\bar{Y}-\bar{c}/\hat{\alpha} \quad (\bar{c}=\sum_{i=1}^{m}c_i/m). \tag{2·109}$$

即使m不甚大，这时（2·108）近似程度较低，但（2·108）两边的差，符号正负都有。故经过最小二乘处理，所得解(2·109)一般仍不差。这当然不是说m的大小无关紧要。只是说，相对于样本分位数法而言，此法对m的要求较低。因为当m很小时，样本分位数作为总体分布G的分位数的估计，误差太大，而导致u，α的估计很不准。此法所受影响则小些。当然，这里有两个前提：一是n已足够大使极值分布可用，二是各组极值$Y_1, \cdots, Y_m$基本上是独立同分布。如有极值概率纸，可将m个点$(Y_{(i)}, \frac{i}{m+1})$，$i=1, \cdots, m$，描在这种纸上，若这散点图基本上是一直线趋势，则上述要求可认为基本满足。作为近似，可用目测法画出一条回归直线，由之定出$\hat{u}$和$\hat{\alpha}$。一般这与用公式(2·109)算得的相去不远。

3．极大似然估计法

此法效率较高(当然也是从大样本观点)，但计算也较繁。先写出$(Y_1, \cdots, Y_m)$的似然函数L：

$$L=\alpha^m \exp(-\sum_{i=1}^{m}e^{-\alpha(Y_i-u)})\exp(-\alpha\sum_{i=1}^{m}(Y_i-u)).$$

此式的来历简单：由Y_i的分布函数$\exp(-e^{-\alpha(x-u)})$对x求导，得Y_i的概率密度函数，再根据似然函数的定义即得。记$h=e^{-\alpha u}$，$Z=\sum_{i=1}^{m}e^{-\alpha Y_i}/m$，得

$$\log L=m(\log\alpha-\alpha(\bar{Y}-u)-Z/h),$$

把此式分别对u和α求偏导数并令之为0，得方程组

$$e^{\alpha u}\sum_{i=1}^{m}e^{-\alpha Y_i}=m,$$
$$\sum_{i=1}^{m}Y_i e^{-\alpha Y_i}/\sum_{i=1}^{m}e^{-\alpha Y_i}+\frac{1}{\alpha}=\bar{Y}.$$

用数值方法，先由第二式解出α，以其结果代入第一式的α而解出u。

有人可能会认为，用下面的方法处理极值问题，既简单直接又可能效率更高：既然n和m都不太小，则mn应是一相当大的数通过mn个样本$X_1,\cdots,X_{mn}$去估计原总体分布F，比方说就用经验分布$\hat{F}$，其精度应较高。然后，Y_i是n个观察值的最大值，其分布应为F^n，可以用$\hat{F}^n$去估计。利用$\hat{F}^n$，就可以解决诸如寻找"百年一遇"界限的问题。这种做法表面上直截明了，实际上不一定可行。一则在应用中，每组内n个观察值同分布的要求不见得很好地满足这影响了F的估计$\hat{F}$的精度，例如，一条河流在夏秋之交每日的水位虽有随机性，但总的看高于冬春之交的枯水期的水位。可是，尽管没有这个同分布性，极(大、小)值的分布却不受多大影响。其次，即使F的估计$\hat{F}$有较好的精度，但n一般较大，$\hat{F}^n$作为F^n的估计，精度就不一定好。最后，在有的问题中，只有极值的记录，原始记录或没有或不全。这时，刚才所描述的方法根本无法使用，但前面介绍的方法（它只用到各组的极值$Y_1,\cdots,Y_n$）则不受影响。

在结束本章之前说几句关于截尾数据的问题，在通常情况下所谓截尾数据，指的是下面的情况：没有能观察到全部的次序样本$X_{(1)}\leqslant\cdots\leqslant X_{(n)}$，而只观察到其一部分$X_{(a)}\leqslant X_{(a+1)}\leqslant\cdots\leqslant X_{(b)}$。例如$n$个元件分别观察其寿命，预定到第$r$个失效时试验停止，则$a=1$而$b=r$，$a$，$b$都为非随机的。也可以先指定一个时间$t$，试验进行到该时刻为止，且只记下当时已失效的元件寿命数据，则或者根本无可记录（但也知道了这n个受测元件的寿命都大于t，这个信息也可用于统计推断），或者是$a=1$而$1\leqslant b\leqslant n$，b是随机的。又如X是某个量在一定仪器上测出的值。而该仪器只能读出界限A、B之间的测定值。这时a，b都是随机的。

处理截尾数据的统计方法很多，多数属于参数统计范围。我

们不打算深入这个课题，只结合本章内容作一个附注：前面所讲方法，大多只涉及少数几个次序统计量，例如极值方法只用到$X_{(n)}$或$X_{(1)}$，估计对称中心只用到样本中位数等，只要所用到的那一个或几个次序统计量属于被记录的范围(即在$X_{(a)},\cdots, X_{(b)}$内)，则以前各节的方法畅通无阻。道理很简单：我们不妨假装认为全体$X_{(1)}, \cdots, X_{(n)}$都在，这当然不影响$X_{(a)},\cdots, X_{(b)}$之值。因此，有时可以有意识地调整方法，使所需的次序统计量都在被记录的范围内。试举例以明之：设变量X, Y分别有分布$F(x)$和$F(x-\theta)$。为估计θ，对X, Y分别拟作m次和n次观察。但前者只观察到$X_{(a)}\leqslant\cdots\leqslant X_{(b)}$，而后者只观察到$Y_{(c)}\leqslant\cdots\leqslant Y_{(d)}$。这时，如能找到这样一个介于0与1之间的数$p$，使$X$样本(大小为$m$)的$p$分位数$m_p(X)$属于$X_{(a)},\cdots, X_{(b)}$内，而$Y$(样本大小为$n$)的$p$分位数$m_p(Y)$属于$Y_{(c)},\cdots, Y_{(d)}$内。则$\theta$可用$m_p(Y)-m_p(X)$去估计之。上述$p$是在有了样本以后才选定的，从最严格的理论观点看，这与p的选定不依赖于样本的情况有所不同，但是，类似于这种做法，在实践中不时见到(如在用x^2法作拟合优度检验时，分组随样本情况而定；在采用何种回归模型的问题中参考散点图等)，一般多不予深究。方便的看法是：就把这个p看成是事先选定的。

习　题

2-1　写出恒等式(2·4)的完整证明。

2-2　证明：设$X_{(1)}\leqslant\cdots\leqslant X_{(n)}$为分布$F$中抽出的简单样本的次序统计量，且对某个$r\leqslant n$，$X_{(r)}$的概率密度存在。则$F$本身的概率密度存在(不熟悉绝对连续概念的读者可略去本题)。

2-3　记号同上题。设$1\leqslant r<s\leqslant n$。试写出$(X_{(r)}, X_{(s)})$的联合分布函数，并对之求导以得出公式(2·13)

2-4　设随机变量X有分布$F(x)$。若F并非处处连续，则$F(X)$不服从(0,1)均匀分布。

2-5　记号同上题。设F处处连续。把（2·55）式定义的函数$G(x)$记为$F^{-1}(x)$（这记号意味着把这样定义的G视为F的反函数，虽则当F不处处严增时，通常意义下的反函数不存在），又U为（0,1）均匀分布，则$X \overset{d}{=} F^{-1}(U)$（此题形式上是上题之逆：上题给出$U \overset{d}{=} F(X)$。两边取$F^{-1}$）。

2-6　设$U_{(1)} \leqslant \cdots \leqslant U_{(n)}$为（0,1）均匀分布的次序样本。记$V_1 = U_{(1)}, V_{(i)} = U_{(i)} - U_{(i-1)}$，$i = 2$，…，$n$，$V_{n+1} = 1 - U_{(n)}$。证明$V_1, \cdots$，$V_{n+1}$同分布但非独立（任意一对不独立）。

2-7　以μ记对称分布F的对称中心。设$X_1, \cdots$，X_n为F的简单样本，$\hat{m}$为其样本中位数。求证$\hat{m}$是μ的无偏估计。此题可直接证（即通过样本中位数之分布）或利用对称性用一个简单技巧证得。

2-8　用两种方法算极差R的期望，证明结果一致：(1)用极差分布（2·25）。(2) 用公式$E(R) = EX_{(n)} - EX_{(1)}$。

2-9　若X不是有界随机变量，则极差R的期望$E(R)$必随$n \to \infty$而趋于无穷。反之，若X有界，则当$n \to \infty$时$E(R) \to \sup X - \inf X$。这里$\operatorname{Sup} X$和$\inf X$分别指X的"实质"上、下确界。例如，$\operatorname{Sup} X$的意义是：$P(X \leqslant \sup X) = 1$，但对任给$\varepsilon > 0$有$P(X \leqslant \sup X - \varepsilon) < 1$。又除非$X$退化，$E(R)$必是$n$的严格增加函数。

2-10　当总体为（0,1）均匀分布且n为偶数时，求出样本中位数的密度。

2-11　用条件分布的方法求$aX_{(r)} + bX_{(s)}$的分布（记号同第1题）。

2-12　证明：任一对称分布的对称中心必唯一。

2-13　举一个简单反例证明：样本中位数不一定是总体中位数的无偏估计。

2-14　（续上题）但是，若总体分布对称，则样本中位数必为对称中心的无偏但不必相合的估计。

2-15　（用引理2·2的记号）对$a \leqslant 0$的情况，完成引理2·2的

证明。

2-16　在样本分位数的一般定义（2·16）之下，证明其渐近正态性（在定理2·2的基础上）。

2-17　证明（2·57）和（2·59）是一回事。

2-18　试从Moore定理（定理2·4）推出 Lindeberg 中心极限定理（iid.情况）。

2-19　利用定理 2·4，写出两边切尾比例不同时，切尾均值的极限定理。

2-20　当底分布为负指数分布或 Cauchy 分布时，通过直接计算去证明Von Mises定理（定理2·3）的结论。

2-21　设 $\mathscr{F}$ 为一切一维分布构成的分布族，$X_1, \cdots, X_n$ 为一维简单样本。记 $T_1=(X_1, \cdots, X_n)$，$T_2=X_1$。证明：T_1 充分非完全，T_2完全非充分。

2-22　$\mathscr{F}$及$X_1, \cdots, X_n$的意义同上题，又设n为偶数 $2m$。以$Y_1 \leqslant \cdots \leqslant Y_m$和$Z_1 \leqslant \cdots \leqslant Z_m$分别记$X_1, \cdots, X_m$及$X_{m+1}, \cdots, X_{2m}$的次序统计量。证明：统计量$(Y_1, \cdots, Y_m, Z_1, \cdots, Z_m)$充分但非完全（此题结果不必 n 为偶数，且两段变量个数任意时也对，证明略繁）。

2-23　设总体分布族为均匀分布族$\{R(\theta_1, \theta_2): -\infty<\theta_1<\theta_2<\infty\}$。证明当样本大小为 2 时，次序统计量$(X_{(1)}, X_{(2)})$为完全统计量。当$n \geqslant 3$时，$(X_{(1)}, \cdots, X_{(n)})$不为完全。

2-24　假定在例2·1中，分布F有密度f，$f(0)>0$且$f(x)$在$x=0$处连续。试在这些假定下，求当$\lambda \uparrow \frac{1}{2}$时，（2·68）式中的$\sigma^2$极限，并解释所得结果。

第三章 U统计量法

§3·1 从统计问题引进U统计量

一、U统计量的定义：一样本情况

在§2·3中，我们指出了，在总体分布族满足很广泛的条件下，简单样本的次序统计量有充分性与完全性（其中充分性对总体分布族不要求任何条件）。这个重要事实与参数估计理论中的Lehmann-Scheffe定理结合，就可以证明某些估计量是最小方差无偏估计。样本均值作为总体均值的估计是一个简单而典型的例子。把这个例子加以引伸，可以在下面的一般模式，从中自然地引出U统计量的定义。

设总体分布F(暂设是一维分布)属于一定的分布族$\mathscr{F}$。设$\mathscr{F}$满足§2·3提出的有关次序统计量完全性的条件。设$\theta(F)$是定义在$\mathscr{F}$上的一个取实数值的泛函——从统计的观点看，$\theta(F)$无非是分布F的某项特征，例如$\theta(F)$可以是F的期望，中位数，变异系数或方差等。习惯上也把这种$\theta(F)$称为分布F的参数，但这不过是指其值由F决定这个事实，与参数统计中那种决定分布形状的实参数不是一回事。

现设从总体F中抽出了大小为n的简单样本$X_1,\cdots,X_n$，要依据它去估计$\theta(F)$，希望找到$\theta(F)$的最小方差无偏估计。一般，只用到少数几个样本的无偏估计比较好找（当然，在有的问题中无偏估计根本不存在，那就是另一回事了）。设想我们找到了只依赖m个样本（$m\leqslant n$）$X_1,\cdots,X_m$的无偏估计$h(X_1,\cdots,X_m)$。例如当$\theta(F)=F$的数学期望时，可取$m=1$而$h(x)=x$。由于$h(X_1,\cdots,X_m)$只用了一小部分样本，它的性能不可能很好。但是，由于$X_1,\cdots,X_n$为i.i.d.对任何固定的、介于

1与 n 之间的、两两不同的自然数 $i_1,\cdots, i_m$, $h(X_{i_1},\cdots,X_{i_m})$ 也是 $\theta(F)$ 的无偏估计。遍取所有这样的 $(i_1,\cdots, i_m)$，我们就得到 $n(n-1)\cdots(n-m+1)$ 个无偏估计。直观上觉得，将它们平均，会得到性能更好的无偏估计：

$$U_n=U_n(X_1,\cdots, X_n)=\frac{1}{n(n-1)\cdots(n-m+1)}\sum\nolimits^{*}{}_{i_1,\cdots,i_m}\cdot h(X_{i_1},\cdots X_{i_m}), \tag{3·1}$$

这里$\sum *$中的 * 号表示求和范围是：$i_1,\cdots, i_m$ 互不相同，且都是不超过 n 的自然数。

U_n 的无偏性是显然的。它还是最小方差的无偏估计。为证此只须注意到：（3·1）式右端的表达式显然不依赖于 $X_1,\cdots, X_n$ 的排列次序，因而只依赖于 $X_1,\cdots, X_n$ 的次序统计量。按关于总体分布族 $\mathscr{F}$ 的假定，次序统计量有充分完全性，因而据 Lehmann-Scheffe 定理，U_n 确是 $\theta(F)$ 的最小方差无偏估计。这引出下述关于U统计量的定义：

定义 3·1 设 $X_1,\cdots,X_n$ 为样本，h 为m个变元的函数，$m\leqslant n$，则由（3·1）式定义的 U_n 称为U统计量。或更仔细地，是以函数 h 为核的，基于样本 $X_1,\cdots,X_n$ 的U统计量。

此定义中并未要求 $X_1,\cdots,X_n$ 为 iid，这是因为，有时需要考虑这样的U统计量，其样本不同分布或不独立，虽在本教程中不会碰到这种情形。

如果 h 为对称函数，则（3·1）式可略简化些（h 对称是指 h 之值与其变元之次序无关）。例如当 $m=3$。在 $h(X_1,X_3,X_4)$, $h(X_1,X_4,X_3)$, $h(X_3,X_1,X_4),\cdots$, $h(X_4,X_3,X_1)$ 等 6 个相同项中，只用保留其一就够了（乘以 6 这个因子）。不难看出，简化后结果是：

$$U_n=\sum\nolimits_{1\leqslant i_1<i_2<\cdots<i_m\leqslant n}h(X_{i_1},X_{i_2},\cdots, X_{i_m})\Big/\binom{n}{m}. \tag{3·2}$$

如果 h 本非对称，则可以先将其对称化。办法是引进下述函数h^*：

$$h^*(x_1,\cdots,x_m)=\sum\nolimits_{i_1\cdots i_m}^{**} h(x_{i_1},\cdots,x_{i_m})/m!,$$

和号中的＊＊表示和范围是：$i_1,\cdots i_m$ 互不相同且都不超过m。换句话说，$(i_1,\cdots,i_m)$取 $1,\cdots,m$ 的一切可能的置换，其数有$m!$个。引进 h^* 后，（3·1）式即可改写为（3·2），只须把其中的 h 改为 h^*，而 h^* 显然是对称的．因此，在以后我们多假定核是对称函数。

读者容易看出：若 $\theta(F)=F$ 的期望而取 $h(x)=x$，则U统计量就是$\bar{X}$。现考虑一个略复杂一点的例子。

例 3·1 以$\mathscr{F}$记一切其方差有限的一维分布族，要找方差的最小方差无偏估计。

设有简单样本 $X_1,\cdots,X_n$。取 $h(x_1,x_2)=x_1^2-x_1x_2$（注意这不是对称核）。则易见

$$E_Fh(X_1,X_2)=E_FX_1^2-E_FX_1\cdot E_FX_2=E_FX_1^2-(E_FX_1)^2$$
$$=F\text{ 的方差}$$

故 h 可取为核．有

$$U_n=\frac{1}{n(n-1)}\sum\nolimits_{i_1,i_2}^{*}(X_{i_1}^2-X_{i_1}X_{i_2}),$$

求和范围为 $1\leqslant i_1,\ i_2\leqslant n,\ i_1\neq i_2$。在$\sum\nolimits_{i_1,i_2}^{*}X_{i_1}^2$中，每个 X_j^2 出现 $n-1$ 次，因每个 j 可与另一个不超过 n 且不等于j 的自然数 k 搭配，而这种k 有 $n-1$ 个．故

$$\sum\nolimits_{i_1,i_2}^{*}X_{i_1}^2=(n-1)\sum_{i=1}^{n}X_i^2.$$

另一方面，

$$\sum\nolimits_{i_1,i_2}^{*}X_{i_1}X_{i_2}=\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}X_{i_1}X_{i_2}-\sum_{i=1}^{n}X_i^2=n^2\bar{X}^2-\sum_{i=1}^{n}X_i^2,$$

结合以上三式，得

$$U_n=\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right)=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2.$$

这就是通常的样本方差．因为按定理 2·7，对本例中的总体分布族$\mathscr{F}$而言，次序统计量有完全性。故样本方差是总体方差的方差

最小无偏估计。

按本例的做法可得到任意阶总体中心矩的最小方差无偏估计。阶数愈高，计算愈繁，但原则上无困难，以 3 阶中心矩 μ_3 为例：

$$\mu_3=E(x-m)^3=EX^3-3mEX^2+2m^3,$$

其一个无偏估计显然是 $X_1^3-3X_1X_2^2+2X_1X_2X_3$。以此作为(3·1)式中的核 $h(x_1,x_2,x_3)$，可算出 U_n，它就是 μ_3 的最小方差无偏估计。当然，对总体分布族 $\mathscr{F}$ 要有一定的假定。例如，$\mathscr{F}$ 是一切其 3 阶矩有限的一维分布族。

同一个量 $\theta(F)$ 的无偏估计不止一个。因此，我们可能找到两个核函数 $h_1(x_1,\cdots,x_{m1})$ 和 $h_2(x_1,\cdots,x_{m2})$，同时满足条件

$$E_F h_i(X_1,\cdots,X_{mi})=\theta(F),\ i=1,2。$$

分别从 h_1,h_2 出发利用(3·1)，就得到两个U统计量，暂记为U_{n1}和U_{n2}。在$\mathscr{F}$满足适当条件时，它们都是 $\theta(F)$ 的最小方差无偏估计。这样一来，岂非最小方差无偏估计可能有很多？其实不然。因按完全性定义，只依赖于某一完全统计量的无偏估计，实质上（就是说，以概率 1 ）只有一个。所以，不论你从什么核出发，最后所得的U统计量形式一样。例如，当 $\theta(F)=F$ 的期望时，你可以取 $h_1(x_1)=x_1$ 或取 $h_2(x_1,x_2)=(x_1+x_2)/2$ 为核，它们引出的U统计量都是样本均值 $\overline{X}$。

这样就产生一个有趣的问题：对给定的 $\theta(F)$，要决定最小的 m，使只含m变元的核存在。这样的m称为 $\theta(F)$的“级”。决定一个泛函 $\theta(F)$的级有时不难。例如，总体期望有只依赖一个样本的无偏估计，故其级当然是 1。由例 3·1 知总体方差的级不超过 2 。不难证明就是 2（习题 2 ）。对某些情况，级的确定是一个很困难的问题。由于此问题与统计应用关系不大，在此不细谈了。

以上我们假定了总体是一维的。若总体为多维，情况也类似。只有一点需注意：在一维情况，次序统计量按大小排列，其意义很

清楚。在多维情况，次序就不清楚，但这关系不大。设 $X_1,\cdots,X_n$ 是从一多维总体分布 F 中抽出的简单样本，因之每个 X_i 均属多维。引进一个统计量 T：

$$T=T(X_1,\cdots,X_n)=\{X_1,\cdots,X_n\}, \tag{3·3}$$

这里括号的意义是指集合。就是说，T 就是由以这 n 个样本为元素构成的集合，但还有一点细微的差别：在通常集合论中，重复的元素只计一次。此处则不然，重复的都要保留。在这样的定义之下，§2·3 关于 T 的充分性的论证可不作任何改变用于此处的 T。关于 T 的完全性对 $\mathscr{F}$ 的要求，也与一维情形一样，这一点不深入了。因此，定义 U 统计量的前提仍适合，而我们仍可循着前面的思路达到定义 3·1。

有的读者可能会对(3·3)这种类型的统计量 T 感到不习惯，因它不取实数或实向量为值，其实，在统计量的一般定义中，主要之点是其"值"（广义的值，如集、函数之类）只取决于样本。是否取实数值并不关紧要、读者还应注意：(3·3)定义的 T，其实质在于要"忘掉"原样本中 n 项的次序(即 X_1 最先，其次 X_2 等等)，这与在一维时把 n 项按大小重排起的作用完全一样。故(3·3)与次序统计量是貌异而实同。

举一个简单例子。

例 3·2 设 F 为二维分布，$\mathscr{F}$ 为所有那些其二阶矩(两个分量的二阶矩)有限的二维分布族。又 (X_{1i},X_{2i})，$i=1,\cdots,n$，为抽自总体 F 的简单样本。要依据它去估计 $\theta(F)$＝总体协方差。

若记 $X_i=(X_{1i},X_{2i})$，则易见 $h(X_1,X_2)=X_{11}X_{21}-X_{11}X_{22}$ 是 $\theta(F)$ 的一个无偏估计。于是，按(3·1)有

$$U_n=\frac{1}{n(n-1)}\sum\nolimits^*_{i_1,i_2}(X_{1i_1}X_{2i_1}-X_{1i_1}X_{2i_2}),$$

求和范围为 $i_1\neq i_2$，$1\leqslant i_1\leqslant n$，$1\leqslant i_2\leqslant n$。经过简单计算，得

$$U_n=\frac{1}{n-1}\sum_{i=1}^{n}(X_{1i}-\bar{X}_1)(X_{2i}-\bar{X}_2),$$

其中 $\bar{X}_j=\sum_{i=1}^{n}X_{ji}/n$。这就是通常的样本协方差。本例的 $\mathscr{F}$ 满足使（3·3）定义的 T 为完全统计量之条件，因此证明了：在所说的总体分布族之下，样本协方差是总体协方差的最小方差无偏估计。

二、U 统计量的定义：两样本情况

在一段中，我们考虑了样本是从一个总体分布 F 中抽出的情况。如在第一章中交代过的，这类问题在统计中统称为"一样本问题"。所谓"多样本问题"，则是指在同一统计问题中涉及多于一组的样本，每组样本系从某一总体中抽出。在这类问题中 U 统计量法也常有用，特别是两组的情况。故我们就这个情况来讨论。

设有两个总体，其分布分别为 F 和 G，假定 F 属于某分布族 $\mathscr{F}$，而 G 属于某分布族 $\mathscr{G}$。设有一个定义于 $\mathscr{F}\times\mathscr{G}$ 上的实值泛函 $\theta(F,G)$。从总体 F 中抽出简单样本 $X_1,\cdots,X_{n_1}$ 而从 G 中抽出简单样本 $Y_1,\cdots,Y_{n_2}$，且设 $X_1,\cdots,X_{n_1}$，$Y_1,\cdots,Y_{n_2}$ 全体独立。要利用这些样本去估计 $\theta(F,G)$。

推理的过程与一样本情况相同。先设法找到一个函数 $h(x_1,\cdots,x_{m_1};\ y_1,\cdots,y_{m_2})$，使 $h(X_1,\cdots,X_{m_1};\ Y_1,\cdots,Y_{m_2})$ 为 $\theta(F,G)$ 的无偏估计。接着，利用每组样本独立同分布，而两组样本独立，知对任何互不相同而不超过 n_1 的 $i_1,i_2,\cdots,i_{m_1}$ 及互不相同而不超过 n_2 的 $j_1,j_2,\cdots,j_{m_2}$，$h(X_{i_1},\cdots,X_{i_{m_1}};\ Y_{j_1},\cdots,Y_{j_{m_2}})$ 也是 $\theta(F,G)$ 的无偏估计。把所有这些加以平均，得到

$$U_{n_1n_2}=U_{n_1n_2}(X_1,\cdots,X_{n_1};\ Y_1,\cdots,Y_{n_2})$$

$$=\frac{1}{n_1(n_1-1)\cdots(n_1-m_1+1)n_2(n_2-1)(n_2-m_2+1)}$$

$$\textstyle\sum^{*}_{i_1\cdots i_{m_1}j_1\cdots j_{m_2}}h(X_{i_1},\cdots,X_{i_{m_1}};Y_{j_1},\cdots,Y_{j_{m_2}}),\quad(3\cdot4)$$

$\sum^*$ 表示求和的范围是满足刚才描述过的那些条件的 $(i_1,\cdots,i_{m_1},j_1,\cdots,j_{m_2})$。

定义 3·2　(3·4)确定的 U_n 称为以 h 为核的、基于两组样

本 $X_1,\cdots,X_{n_1}$ 和 $Y_1,\cdots,Y_{n_2}$ 的U统计量。

与一样本的情况一样，可以证明在一定条件下，这样定义的 U_n 是 $\theta(F,G)$的最小方差无偏估计，这条件就是：$\mathscr{F}$ 和$\mathscr{G}$ 这两个分布族都满足定理 2·7 中施加在分布族$\mathscr{F}$上的条件。细节这里不涉及了。

当 h 分别对两组变量为对称时，也就是说，当 $y_1,\cdots,y_{m_2}$ 固定时 h 是 $x_1,\cdots,x_{m_1}$ 的对称函数，而当 $x_1,\cdots,x_{m_1}$ 固定时，h 为 $y_1,\cdots,y_{m_2}$ 的对称函数，则（3·4）式可简化为

$$U_{n_1n_2}=\sum_{\substack{1\leqslant i_1<\cdots<i_{m_1}\leqslant n_1\\ 1\leqslant j_1<\cdots<j_{m_2}\leqslant n_2}} h(X_{i_1},\cdots,X_{i_{m_1}};\ Y_{j_1},\cdots,Y_{j_{m_2}})\Big/\binom{n_1}{m_1}\binom{n_2}{m_2} \qquad (3\cdot5)$$

当 h 不为对称时，也可以先将其对称化，即用函数

$h^*(x_1,\cdots,x_{m_1};\ y_1,\cdots,y_{m_2})=\sum^{**}h(x_{i_1},\cdots,x_{i_{m_1}};\ y_{j_1},\cdots,y_{j_{m_2}})/m_1!m_2!$ 去代替 h。$\sum^{**}$ 表示求和范围为：$(i_1,\cdots,i_{m_1})$ 是 $1,\cdots,m_1$的一切置换，而$(j_1,\cdots,j_{m_2})$ 是 $1,\cdots,m_2$ 的一切置换。

例 3·3 把变量X设想为一种产品在一定工艺规程之下的质量指标，指标值愈大产品质量愈好。以Y记这同一产品在一种经过改进的工艺规程之下的质量指标。如预先的设想正确，则工艺上的改变应有助于提高质量指标。在统计上反映这一点的一个方法是：应有 $P(X<Y)>1/2$，而如所作改变无助于提高质量，则应有 $P(X<Y)=1/2$。此处我们假定 X，Y 的分布 F，G 都处处连续，且X，Y独立。值得注意的是X，Y的意义如何理解。由于同一产品不可能既是在原工艺又是在新工艺下制造的，X，Y并不是在同一件产品上量出的两个值，而应这样去理解：从原工艺下生产的产品中随机抽取一个，量得其指标为X；又独立地从新工艺下生产的产品中随机抽取一个，量得其指标为Y。

新工艺即使比原工艺有所改进，也不能保证，在新工艺下生产的每一件产品，其指标必高于在原工艺下生产的每一件产品。

而只能说：新工艺"倾向于"生产出指标较高的产品。其确切含义可解释如下：指定任一常数 x；把质量分成两类：一类是 $\leqslant x$，一类是 $>x$。相对于后者，前一类属于质量较差的类。在原工艺下这一类产品的概率为 $P(X\leqslant x)=F(x)$，而在新工艺下，其概率则为 $P(Y\leqslant x)=G(x)$。如果

$$G(x)\leqslant F(x) \text{对一切实数} x\text{，且} F\neq G, \tag{3·6}$$

则在新工艺下，产出质量较差的产品的概率，总不超过（且有时确小于）其在原工艺下的概率。在这个意义下（注意这是严格的数学意义），我们说新工艺下产品质量优于原工艺。

定义 3·3 若 X，Y 为两个一维随机变量，其分布函数分别为 F 和 G。若（3·6）成立，则称 Y 随机地大于 X，有时记为 $Y\overset{r}{>}X$。

如果分布 F、G 都处处连续且 X，Y 独立，则有

$$P(X<Y)=\int_{-\infty}^{\infty}F(x)dG(x) \tag{3·7}$$

（3·7）易用条件概率方法证明：固定 $Y=x$。由于 X，Y 独立且分布 F 连续，有

$$P(X<Y|Y=x)=P(X<x|Y=x)=F(x)$$

再注意 Y 有分布 G，即得（3·7）。当 X，Y 同分布时有

$$\int_{-\infty}^{\infty}F(x)dG(x)=\int_{-\infty}^{\infty}F(x)dF(x)=\int_0^1 tdt=1/2,$$

而当 $Y\overset{r}{>}X$ 时，$F(x)\geqslant G(x)$，（3·7）式右边将 $\geqslant 1/2$。因为 $F\neq G$，等号不成立（严格证明留给读者），故将有 $P(Y>X)>1/2$。

$Y\overset{r}{>}X$ 的一个最重要的例子是 $Y\overset{d}{=}X+\theta$ 而 $\theta>0$。建议读者写出仔细证明。

设现在我们分别从原工艺和新工艺下抽出其 n_1 个和 n_2 个产品，量得其质量指标分别为 $X_1,\cdots,X_{n_1}$ 及 $Y_1,\cdots,Y_{n_2}$。要依据

它来检验假设

H：两种工艺下产品质量一样，即 $F=G$，

其对立假设为

K：新工艺下产品质量较优，即（3·6）成立。

根据前面的讨论，我们可以把

$$\theta=\theta(F,G)=P_{F,G}(Y>X)=P(Y>X)$$

作为一个检验的基准。也就是说，我们努力通过样本给 θ 一个良好的估计。若此估计接近 1/2，则我们无充分理由否定原假设 H。反之，若此估计显著大于 1/2，则将否定 H。为估计 θ 用得着 U 统计量的方法：令

$$h(x_1;\ y_1)=\begin{cases}1, & \text{当 } x_1<y_1\\ 0, & \text{其他}\end{cases} \tag{3·8}$$

则有 $Eh(X_1;\ Y_1)=\theta$。于是按定义 3·2，以它为核而产生的 U 统计量

$$U_{n1n2}=\sum_{i=1}^{n1}\ \sum_{j=1}^{n2} I(X_i<Y_j)/_{n1n2} \tag{3·9}$$

是 θ 的最小方差无偏估计。

为了计算（3·9）式右边的分子，我们把 n_1+n_2 个样本按由小到大排序。由于分布连续，可以假定这 n_1+n_2 个样本互不相同。以 R_i 记 Y_i 在这个排列中的位次（最小者位次为 1，其次为 2，余类推）。R_i 称为 Y_i（在这 n_1+n_2 个样本中的）"秩"。于是，合样本中有 R_i-1 个小于 Y_i，设其中有 c_i 个 Y 样本，则 X 样本有 $R_i-(c_i+1)$ 个。这些 X 样本与 Y_i 配对，在（3·9）式分子中产生 $R_i-(c_i+1)$ 个 1。因此

$$(3·9)\text{式的分子}=\sum_{i=1}^{n2}(R_i-c_i-1)。$$

因为 $\sum_{i=1}^{n2}(c_i+1)=1+2+\cdots+n_2=n_2(n_2+1)/2$，故

$$U_{n_1n_2}=\Big(\sum_{i=1}^{n2}R_i-n_2(n_2+1)/2\Big)/n_1n_2。 \tag{3·10}$$

按前面的讨论，得出如下的检验法：当 $U_{n_1n_2}$ 超过某常数C时，否定原假设H，不然就接受H。这个检验叫做 Mann-Whitney 检验，是这两位学者在 1947 年提出的。文献中把 $n_1n_2U_{n_1n_2}$ 叫做 Mann-Whitney 统计量。在 1947 年时尚未提出U统计量的概念尔后明确了它是U统计量的一个简单例子，就得以方便地证明其一些深入的大样本性质。

从（3·10）式看出：Mann-Whitney 检验与下述检验等价：记 $R=\sum_{i=1}^{n_2}R_i$。当R大于某常数C时，否定原假设H，不然就接受H。这个检验比 Mann-Whitney 检验更早，它是 Wilcoxon 在 1945 年提出的，在文献中称为 Wilcoxon 两样本秩和检验，因为R是Y样本之秩之和。这个检验属于下一章中要仔细讨论的秩检验的范围。人们也常把这个检验称为 Wilcoxon-Mann-Whitney 检验。

与用于估计问题相比，U统计量用于检验问题稍有其不同之处。在估计问题中，$\theta(F)$（或 $\theta(F,G)$）是早有的。在检验问题中，开始并无 θ，而要求找出这样一个 θ：(1)其值在原假设成立时是明确的（如 $\theta=\theta_0$ 或$\theta\leqslant\theta_0$ 之类）.(2)当偏离原假设时，θ 之值能“敏感地”反映这一点。(3)容许用U统计量法去处理。这样的 θ 有时无法找到，有时可以有很多，其优劣不易在直观上判出。例如，在本例中若假定Y有分布 $G(x)=F(x-\Delta),\Delta\geqslant0$ 为未知参数。则 Wilcoxon 检验可用，但也可以取 θ 就等于Δ。它可以（再进一步假定F有有限的数学期望）通过核 $h(X_1;Y_1)=Y_1-X_1$ 用U统计量法去处理，结果将得出下述检验：当 $\bar{Y}-\bar{X}$ 大时否定原假设。这里界限的确定比上例要复杂些，因为$\bar{Y}-\bar{X}$ 在原假下并非分布无关。可是单从这两个检验的内容看，直观上分不出优劣。这当然取决于其他条件。在下一章中我们会涉及这个问题，现在再考察一个较为复杂的例子。

例 3·4 再考察我们曾多次提到过的两样本问题：简单样本

$X_1, \cdots, X_{n_1}$ 和 $Y_1, \cdots, Y_{n_2}$ 分别抽自一维总体分布 F 和 G. 只假定 F 和 G 连续，其他一概未知. 要检验 $H: F=G$ 这个原假设. 本例与例 3·3 不一样之处在于此处未假定对立假设有"方向性". 因此，如仍以 $P(X<Y)$ 作指标就不恰当. 因为，当 H 成立时固然有 $P(X<Y)=1/2$，但 H 不成立时，此概率也可以是 1/2. 因之这概率之值不构成分辨原假设和对立假设的一个指标.

在例 1·1 中我们曾指出本问题的一个检验法——Смирнов 检验. 这检验有"全方位"性，是一个合适的检验. 但它不能用 U 统计量的方法去处理.

为要用 U 统计量方法去处理这个问题，必须找到这样一个 $\theta(F, G)$. 它能反映原假设和对立假设的差距，且又有一个简单的无偏估计. 这种 $\theta(F,G)$ 被 Lehmann 找到了，它是

$$\theta(F,G)=\int_{-\infty}^{\infty}[F(x)-G(x)]^2 d(F(x)+G(x)),$$

这样子的 θ 能反映 F 与 G 的差距是明显的：若 $F=G$，则 $\theta(F,G)=0$，否则 $\theta(F,G)>0$. 且一般说，F 与 G 差别愈大，$(F(x)-G(x))^2$ 也愈大，因之 $\theta(F,G)$ 一般也愈大. 故上述第一个要求满足了. 下一步是要找到 θ 的一个无偏估计. 我们来证明：

$$\begin{aligned}\hat{\theta}&=h(X_1,X_2;\ Y_1,Y_2)\\&=-\frac{1}{3}+I(\max(X_1,X_2)<\min(Y_1,Y_2))\\&\quad+I(\max(Y_1,Y_2)<\min(X_1,X_2)),\end{aligned}$$

就是这样一个估计. 事实上，注意到 $\max(X_1, X_2)$, $\min(X_1, X_2)$, $\max(Y_1,Y_2)$, $\min(Y_1,Y_2)$ 分别有分布函数 $F^2(x)$, $1-(1-F(x))^2$, $G^2(x)$, $1-(1-G(x))^2$，得

$$\begin{aligned}E(\hat{\theta}+1/3)=&\int_{-\infty}^{\infty}[1-G(x)]^2 dF^2(x)\\&+\int_{-\infty}^{\infty}[1-F(x)]^2 dG^2(x),\end{aligned}$$

把 $[1-G(x)]^2$ 等展开，逐项积分，使用分部积分并注意 $dF^2=$

$2FdF$ 等（这只在F连续时对），得

$$E(\hat{\theta}+1/3)=2+\int_{-\infty}^{\infty}G^2dF^2(x)+\int_{-\infty}^{\infty}F^2dG^2(x)$$

$$-4\int_{-\infty}^{\infty}F(x)G(x)d(F(x)+G(x))$$

$$=2+\int_{-\infty}^{\infty}d(F^2(x)G^2(x))$$

$$-\int_{-\infty}^{\infty}(F(x)+G(x))^2d(F(x)+G(x))$$

$$+\int_{-\infty}^{\infty}[F(x)-G(x)]^2d(F(x)+G(x))$$

$$=2+1-8/3+\theta(F,G)=1/3+\theta(F,G),$$

这证明了 $E\hat{\theta}=\theta$。于是，以 h 为核，据样本 $X_1,\cdots X_{n_1}$ 和 $Y_1,\cdots,Y_{n_2}$ 作出U统计量 $U_{n_1n_2}$。然后，当$|U_{n_1n_2}|$超过某界限时，否定原假设H。界限的确定要依据检验水平 α，并用到 $U_{n_1n_2}$ 的渐近正态性(见§3·2)。

三、U 统计量的方差

我们只仔细讨论一样本U统计量的情况，因为两样本以至多样本情况在方法上无实质差异。

设 $h(x_1,\cdots,x_m)$ 为对称核，而 U_n 为以 h 为核的、基于简单样本 $X_1,\cdots,X_n$ 的U统计量，记 $Eh(X_1,\cdots,X_m)=\theta$。不失普遍性设 $\theta=0$，不然只须以 $h-\theta$ 代 h。对 $c=1,\cdots,m$。令

$$h_c(x_1,\cdots,x_c)=E\{h(X_1,\cdots,X_m)|X_1=x_1,\cdots,X_c=x_c\}$$
$$=E\{h(x_1,\cdots,x_c,X_{c+1},\cdots,X_m)\} \qquad (3\cdot11)$$

这里 $x_1,\cdots,x_c$ 视为常数，而 $X_{c+1},\cdots,X_m$ 则为随机变量。又 h_m 就是 h。记

$$\sigma_c^2=\mathrm{Var}(h_c(X_1,\cdots,X_c)),\quad c=1,\cdots,m$$

容易看出：若假定 $h(X_1,\cdots,X_m)$ 的方差有限，则 $\sigma_c^2<\infty$对 $c=1,\cdots,m$。事实上(注意已假定 $\theta=0$)

$$Eh_c(X_1,\cdots,X_c)=E\{E\{h(X_1,\cdots,X_m)|X_1,\cdots,X_c\}\}$$

$$=Eh(X_1,\cdots,X_m)=0,$$

故 $\mathrm{Var}(h_c(X_1,\cdots,X_c))=Eh_c^2(X_1,\cdots,X_c)$。由（3·11）有

$$\sigma_c^2=Eh_c^2(X_1,\cdots,X_c)\leqslant E\{E\{h^2(X_1,\cdots,X_m)\,|\,X_1,\cdots,X_c\}\}$$

$$=Eh^2(X_1,\cdots,X_m)=\mathrm{Var}h(X_1,\cdots X_m)<\infty,$$

现有

$$\mathrm{Var}\left(\binom{n}{m}U_n\right)=E\left(\binom{n}{m}U_n\right)^2$$

$$=\sum_{\substack{1\leqslant i_1<\cdots<i_m\leqslant n\\ 1\leqslant j_1<\cdots<j_m\leqslant n}}E\{h(X_{i_1},\cdots,X_{i_m})\cdot h(X_{j_1},\cdots,X_{j_m})\}.$$

我们把两集合$\{i_1,\cdots,i_m\}$和$\{j_1,\cdots,j_m\}$的公共元个数记为c。若$c=0$，则由独立性及$\theta=0$之假定，有

$$E(h(X_{i_1},\cdots,X_{i_m})h(X_{j_1},\cdots,X_{j_m}))$$

$$=Eh(X_{i_1},\cdots,X_{i_m})Eh(X_{j_1},\cdots,X_{j_m})=0\cdot 0=0.$$

若$c\neq 0$，则因h为对称函数，不失普遍性可假定公共元即为1，2，$\cdots$，c。这时有

$$E(h(X_{i_1},\cdots,X_{i_m})h(X_{j_1},\cdots,X_{j_m}))$$

$$=E\{E\{h(X_{i_1},\cdots,X_{i_m})h(X_{j_1},\cdots,X_{j_m})\,|\,X_1,\cdots,X_c\}\}$$

$$=E\{E(h(X_{i_1},\cdots,X_{i_m})\,|\,X_1,\cdots,X_c)\cdot E(h(X_{j_1},\cdots,X_{j_m})\,|\,X_1,\cdots,X_c)\}$$

$$=E(h_c^2(X_1,\cdots,X_c))=\mathrm{Var}(h_c(X_1,\cdots,X_c))=\sigma_c^2,$$

而这样的项一共有$\binom{n}{m}\binom{m}{c}\binom{n-m}{m-c}$个。事实上，从$1,2,\cdots,n$中挑选$m$个足标$i_1,\cdots,i_m$，有$\binom{n}{m}$种挑法。$i_1,\cdots,i_m$挑定后，从中挑出那$c$个与$\{j_1,\cdots,j_m\}$公共的有$\binom{m}{c}$个挑法。到此为止，$j_1,\cdots,j_m$已定下了$c$个，剩下的$m-c$个，必须从$i_1,\cdots,i_m$以外那$n-m$个足标中去挑，挑法为$\binom{n-m}{m-c}$。故得上述结果。由此可知，

$$\mathrm{Var}\left(\binom{n}{m} U_n\right) = \sum_{c=1}^{m} \binom{n}{m} \binom{m}{c} \binom{n-m}{m-c} \sigma_c^2,$$

而

$$\mathrm{Var}(U_n) = \binom{n}{m}^{-1} \sum_{c=1}^{m} \binom{m}{c} \binom{n-m}{m-c} \sigma_c^2. \qquad (3\cdot12)$$

两样本情况与此类似，但细节上略繁复一些。这里我们只将结果写下，而建议读者自己把细节补出来。设 $U_{n_1n_2}$ 由（3·5）定义。对 $0 \leqslant c \leqslant m_1, 0 \leqslant d \leqslant m_2$，引进函数

$$\begin{aligned} h_{cd}(x_1, \cdots x_c;\ y_1, \cdots, y_d) = E\{E(h(X_1, \cdots, X_{m_1}; & \\ Y_1, \cdots, Y_{m_2}) \mid X_1 & \\ = x_1, \cdots, X_c = x_c; & \\ Y_1 = y_1, \cdots, Y_d = y_d)\} & \end{aligned}$$

$$= E\{h(x_1, \cdots, x_c, X_{c+1}, \cdots, X_{m_1}; y_1, \cdots, y_d, Y_{d+1}, \cdots, Y_{m_2})\},$$

$\sigma_{cd}^2 = \mathrm{Var}(h_{cd}(X_1 \cdots, X_c;\ Y_1, \cdots, Y_d))$，注意 $\sigma_{00}^2 = 0$，

则

$$\mathrm{Var}(U_{n_1n_2}) = \binom{n_1}{m_1}^{-1} \binom{n_2}{m_2}^{-1} \sum_{c=0}^{m1} \sum_{d=0}^{m2} \binom{m_1}{c} \binom{n_1 - m_1}{m_1 - c} \binom{m_2}{d} \binom{n_2 - m_2}{m_2 - d} \sigma_{cd}^2. \qquad (3\cdot13)$$

值得注意的是：（3·12）和（3·13）都是以对称核为出发点的。当核非对称时应先将其对称化。

例 3·5 据例 3·1，样本方差为 U 统计量。其核经对称化后，为 $h(x_1, x_2) = (x_1 - x_2)^2/2$。仍设总体分布 F 之期望为 0，其 k 阶中心矩记为 μ_k，设 $\mu_4 < \infty$，有

$$h_1(x_1) = E\{(x_1 - X_2)^2/2\} = \frac{1}{2}(x_1^2 + \mu_2),$$

$$\sigma_1^2 = \mathrm{Var}(h_1(X_1)) = \frac{1}{4}\mathrm{Var}(X_1^2) = \frac{1}{4}(\mu_4 - \mu_2^2),$$

$$\begin{aligned} \sigma_2^2 &= \mathrm{Var}\{(X_1 - X_2)^2/2\} \\ &= \frac{1}{4}(E(X_1 - X_2)^4 - [E(X_1 - X_2)^2]^2) \end{aligned}$$

$$=\frac{1}{2}(\mu_4+\mu_2^2),$$

以此代入（3·12），得

$$\operatorname{Var}\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right)=\frac{1}{n}\mu_4-\frac{n-3}{n(n-1)}\mu_2^2. \quad (3\cdot14)$$

例 3·6 考虑（3·9）式所定义的两样本U统计量，此处$m_1=m_2=1$。因X，Y分别有分布F，G，且F，G处处连续，又$h(x_1;\ y_1)=I(x_1<y_1)$。于是有

$$h_{10}(x_1)=1-G(x_1),\quad \sigma_{10}^2=\operatorname{Var}(G(X_1))$$

$$=\int_{-\infty}^{\infty}G^2(x)dF(x)-\left(\int_{-\infty}^{\infty}G(x)dF(x)\right)^2,$$

$$h_{01}(y_1)=F(y_1),\qquad \sigma_{01}^2=\operatorname{Var}(F(Y_1))$$

$$=\int_{-\infty}^{\infty}F^2(x)dG(x)-\left(\int_{-\infty}^{\infty}F(x)dG(x)\right)^2,$$

$$\sigma_{11}^2=\operatorname{Var}(I(X_1<Y_1))$$

$$=\int_{-\infty}^{\infty}F(x)dG(x)-\left(\int_{-\infty}^{\infty}F(x)dG(x)\right)^2.$$

据（3·13），对本例有

$$\operatorname{Var}(U_{n_1n_2})=\frac{1}{n_1n_2}((n_2-1)\sigma_{10}^2+(n_1-1)\sigma_{01}^2+\sigma_{11}^2). \quad (3\cdot15)$$

将前面求得的σ_{10}^2、σ_{01}^2和σ_{11}^2的表达式代入此式，即得$U_{n_1n_2}$的方差表达式，其值依赖于分布F和G，在原假设$F=G$成立的特例下（还要用到分布的连续性），有

$$\int_{-\infty}^{\infty}F(x)\,dG(x)=\int_{-\infty}^{\infty}G(x)\,dF(x)=\int_0^1 tdt=1/2,$$

$$\int_{-\infty}^{\infty}F^2(x)dG(x)=\int_{-\infty}^{\infty}G^2(x)dF(x)=\int_0^1 t^2dt=1/3,$$

代入（3·15），可知在这个特殊情况下有

$$\operatorname{Var}(U_{n_1n_2})=(n_1+n_2+1)/12n_1n_2 \quad (\text{当}F=G), \quad (3\cdot16)$$

此表达式不依赖于F。这一点是秩统计量的共性，见§4·1。

四、U统计量的相合性

设 U_n 是以 $h(x_1,\cdots,x_m)$ 为（对称）核的、基于简单样本 $X_1,\cdots,X_n$ 的 U 统计量。设 $h(X_1,\cdots,X_m)$ 为 $\theta=\theta(F)$ 的无偏估计，则 U_n 也是 θ 的无偏估计。由方差的表达式（3·12）易知：U_n 作为 θ 的估计还是相合的。事实上，由（3·12）易得出 $\lim\limits_{n\to\infty} n\mathrm{Var}(U_n)=m^2\sigma_1^2$，因而 $\lim\limits_{n\to\infty}\mathrm{Var}(U_n)=0$，即 $\lim\limits_{n\to\infty} E(U_n-\theta)^2=0$。这表明当 $n\to\infty$ 时，U_n 依二阶矩收敛（常称为均方收敛）于 θ。即 U_n 是"二阶矩相合"或"均方相合"的，当然更有弱相合性，即 $U_n\xrightarrow{P}\theta$ 当 $n\to\infty$ 时。在下一节中我们将要证明：当 $n\to\infty$ 时，U_n 有渐近正态性。U_n 的弱相合性是这个结果的一个简单推论，但均方相合性不能从渐近正态性推出来。

在两样本的情况，有方差表达式（3·13）。由核表达式容易看出：当

$$n_1\to\infty,\quad n_2\to\infty$$

时，有

$$\mathrm{Var}(U_{n_1n_2})\to 0,$$

于是证明了 $U_{n_1n_2}$ 的均方相合性与弱相合性。若 n_1、n_2 中有一个不趋于∞，则这一点不成立。

§3·2 U统计量的渐近正态性及其应用

U 统计量是美国统计学家 Hoeffding 于1948年在一篇论文中提出的。在该文中 Hoeffding 证明了 U 统计量的渐近正态性。由于有了这个良好性质，U 统计量才能更方便地用于种种统计问题。

一、一样本情况

定理3·1 设 $h(x_1,\cdots,x_m)$ 为对称函数，U_n 为以 h 为核的、基于简单样本 $X_1,\cdots,X$ 的 U 统计量。设

$$Eh^2(X_1,\cdots,X_m)<\infty,\quad \sigma_1^2>0, \tag{3·17}$$

其中 $\delta_1^2=\mathrm{Var}(h_1(X_1))$，而函数 $h_1(x_1)$ 由（3·11）式定义。则当 $n\to\infty$ 时有

$$\sqrt{n}\,(U_n-\theta)\xrightarrow{\mathscr{L}}N(0,m^2\sigma_1^2),\tag{3·18}$$

而 $\theta=E(h(x_1,\cdots,X_m))$。

在证明本定理前对其形式作些解释。首先，因 $E(U_n)=\theta$，$U_n-\theta$ 就是中心化。前面乘数 $\sqrt{n}$ 的得来，就要考虑 U_n 的方差。由（3·12）式易知，当 m 固定而 $n\to\infty$ 时，$\mathrm{Var}(U_n)$ 为 $1/n$ 的数量级，故乘以因子 $\sqrt{n}$，至于极限分布之方差 $m^2\sigma_1^2$，也不难看出。事实上，依（3·12），有

$$\lim_{n\to\infty}\mathrm{Var}(\sqrt{n}\,U_n)=m^2\sigma_1^2.$$

对方差表达式（3·12）细加分析，也指示了证明定理 3·1 的方法。事实上，考虑（3·12）可知，U_n 的方差包含 σ_1^2 一项，其数量级为 $O\left(\frac{1}{n}\right)$，而包含 σ_c^2 的项，当 $c\geqslant 2$ 者，其数量级为 $O\left(\frac{1}{n^c}\right)$，相对于 $O\left(\frac{1}{n}\right)$ 而言都是高阶无穷小。而 σ_1^2 是来自于函数 h_1（见3·11）式。由此启发我们：表达式 $\sum_{i=1}^{n}h_1(X_i)$（经过适当规则化）构成 U_n 的主要部分。而这表达式作为独立同分布和，按 Lindeberg 中心极限定理，依分布收敛于正态分布。

下面转到定理 3·1 的证明，即具体实现这个想法。为此，仍不失普遍性令 $\theta=0$。令

$$W_n=\sqrt{n}\,U_n,\quad V_n=\frac{m}{\sqrt{n}}\sum_{i=1}^{n}h_1(X_i)$$

而来计算 $c_n=E(W_n-V_n)^2$。有

$$c_n=n\mathrm{Var}(U_n)+\mathrm{Var}(V_n)-2E(W_nV_n)$$

前已指出，$\lim_{n\to\infty}n\mathrm{Var}(U_n)=m^2\sigma_1^2$。又 $\mathrm{Var}(V_n)=m^2\mathrm{Var}(h_1(X_1))=m^2\sigma_1^2$，以及

$$E(W_nV_n)=m\cdot E\left(\sum_{i=1}^{n} U_n h_1(X_i)\right)=mnE(U_n h_1(X_1)),$$

$$(3\cdot19)$$

考虑表达式 $E(h(X_{i1},\cdots,X_{im})h_1(X_1))$。注意 $\theta=0$。若 $i_1=1$，则此项为

$$E\{E(h(X_{i1},\cdots,X_{im})h_1(X_1)|X_1)\}=Eh_1^2(X_1)=\sigma_1^2,$$

这种项的数目为$\binom{n-1}{m-1}$。若$i_1>1$，则该项为 0。由此可知

$$E(U_nh_1(x_1))=\binom{n}{m}^{-1}\binom{n-1}{m-1}\sigma_1^2=\frac{m}{n}\sigma_1^2,$$

以此代入（3·19）得 $E(W_nV_n)=m^2\sigma_1^2$。综合上述事实，得 $\lim\limits_{n\to\infty} E(W_n-V_n)^2=0$。于是 W_n 应与 V_n 有相同的极限分布。而据 Lindeberg 定理，V_n 有极限分布 $N(0,\ m^2\sigma_1^2)$。于是证明了定理 3·1。

本定理的证明是 §2·2 提到的（见(2·27) 式下面一段说明）一个一般原则的又一具体使用，即为证明某量有渐近正态性，设法把该量分解为两部分之和，其一部分为一些独立随机变量之和，其渐近正态性由独立和的中心极限定理去处理；另一部分为余项，它在概率上是无穷小，可以忽略不计，或更确切地说，这一项当样本大小 n 趋于无穷时，依概率收敛于 0。

註 1．在本定理条件下显然也成立

$$(U_n-\theta)/\sqrt{\mathrm{Var}(U_n)}\xrightarrow{\mathscr{L}}N(0,1),\qquad(3\cdot20)$$

而且，由于 $(U_n-\theta)/\sqrt{\mathrm{Var}(U_n)}$ 有方差 1（等于 $N(0,1)$ 的方差），一般说来，对同一个 n，它的分布比起变量 $\sqrt{n}\ (U_n-\theta)/(m\sigma_1)$ 的分布（其方差小于 $N(0,1)$ 之方差）来，要更接近 $N(0,1)$一些。在不少应用例子中，至少在原假设成立之下，$\mathrm{Var}(U_n)$可算出且与总体分布无关——只要总体分布属于原假设，在这种情况之下，当然可以而且应该直接使用（3·20）而不必拘泥于 Hoeffding 定理的形式(3·18)。(3·18)有这样一个优点：

如果 $\mathrm{Var}(U_n)$在原假设下并非"分布无关"，则无论用（3·18）还是（3·20），其中之 σ_1^2 或 $\mathrm{Var}(U_n)$ 都须通过样本去估计，但估计 σ_1^2 比估计 $\mathrm{Var}(U_n)$要容易些。

註 2. 在 §3·1 的四中，我们曾在 $E(h^2(X_1, \cdots, X_m))<\infty$ 的条件下（h 为核。注意这条件保证了 U_n 的方差有限），证明了 U_n 作为 θ 的估计，为均方相合及弱相合。利用定理 3·1 的证法，在这同一条件下很容易得出：$U_n \to \theta$，a.s.，即 U_n 也是 θ 的强相合估计。证明梗概如下：由定理 3·1 证明中的计算易看出：

$$E\left(U_n - \frac{m}{n}\sum_{i=1}^{n} h_1(X_i)\right)^2 = \frac{1}{n} E(W_n - V_n)^2 = O\left(\frac{1}{n^2}\right),$$

W_n, V_n 是定理 3·1 证明过程中引入的记号。由上式，用 Чебищев 定理，易知对任给 $\varepsilon>0$ 有

$$\sum_{n=1}^{\infty} P\left(\left|U_n - \frac{m}{n}\sum_{i=1}^{n} h_1(X_i)\right| \geqslant \varepsilon\right) < \infty$$

于是知

$$\lim_{n\to\infty}\left(U_n - \frac{m}{n}\sum_{i=1}^{n} h_1(X_i)\right) = 0, \text{ a.s.} \qquad (3\cdot21)$$

按 Колмогоров 强大数律，且 $Eh_1(X_1)=\theta=0$（已假定 θ 为 0），知 $\frac{m}{n}\sum_{i=1}^{n} h_1(X_i) \to 0$，a.s. 于是 $U_n \to 0$，a.s.

若一开始不假定 $\theta=0$，则（3·21）要用

$$\lim_{n\to\infty}\left\{(U_n - \theta) - \frac{m}{n}\sum_{i=1}^{n} (h_1(X_i) - \theta)\right\} = 0, \text{ a.s.}$$

去取代。由 Колмогоров 定理有 $\sum_{i=1}^{n} (h_1(X_i) - \theta)/n \to 0$，a.s. 故仍得 $U_n \to \theta$，a.s.

从理论的观点看，这个结果有一个不足之处，即假定了核 $h(X_1, \cdots, X_m)$的二阶矩有限，而在 Колмогоров 定理中只假定了一阶矩有限。是否可以只在 $E|h(X_1, \cdots, X_m)|<\infty$的条件下证明 U_n 的强相合性呢？Hoeffding 在 1961 年首先肯定地回答了这个问题。1966 年 Berk发现了U统计量与鞅的关系，因而能很

简单地推出这个事实。这些与统计应用关系不大，故细节在此从略了。

二、两样本情况

在两样本情况下，U统计量渐近正态定理的证明方法，与一样本情况完全类似，故此处只给出定理的陈述，证明细节从略了。读者如有兴趣，应当毫无困难地把它补出来。

仍以 $U_{n_1n_2}$ 记以 $h(x_1,\cdots,x_{m_1};\ y_1,\cdots,y_{m_2})$ 为（对称）核、基于简单样本 $X_1,\cdots,X_{n_1}$ 及 $Y_1,\cdots,Y_{n_2}$ 的U统计量，$\theta=Eh(X_1,\cdots,X_{m_1};\ Y_1,\cdots,Y_{m_2})=EU_{n_1n_2}$。$\sigma_{cd}^2$ 的意义同§3·1，(三)。

定理3·2 设

$$Eh^2(X_1,\cdots,X_{m_1};\ Y_1,\cdots,Y_{m_2})<\infty,\sigma_{10}^2>0,\ \sigma_{01}^2>0,$$

又记 $N=n_1+n_2$，及

$$\sigma_{n_1n_2}^2=N\left(\frac{m_1^2}{n_1}\sigma_{10}^2+\frac{m_2^2}{n_2}\sigma_{01}^2\right), \tag{3·22}$$

则当 $n_1\to\infty$，$n_2\to\infty$时，有

$$\sqrt{N}(U_{n_1n_2}-\theta)/\sigma_{n_1n_2}\xrightarrow{\mathscr{L}}N(0,1)。 \tag{3·23}$$

定理3·1的註1在此当然也适用。即在本定理条件下也有

$$(U_{n_1n_2}-\theta)/\sqrt{\mathrm{Var}(U_{n_1n_2})}\xrightarrow{\mathscr{L}}N(0,1)。 \tag{3·24}$$

三、应用

把§3·2的结果用于大样本统计推断，其原则是简单的。以两样本情况（这个情况在统计应用中比一样本情况要更常见些，尤其在检验问题中）为例说明一下。

1. 假设检验。设 $\delta_{n_1n_2}^2=\mathrm{Var}(U_{n_1n_2})$在原假设下为“分布无关”，即为一已知常数，这时当原假设H成立时，有$(U_{n_1n_2}-\theta_0)/\delta_{n_1n_2}\sim N(0,1)$，此处 θ_0 为 θ 在H成立之下的值（或在H的边缘处之值，见例3·7）这里也就假定了：θ 在 H成立时也须为“分布无关”。又此处记号～是指分布接近。有了这个关系，若原假设H为：$\theta=\theta_0$，则水平 α 的（大样本）否定域可取为 $|U_{n_1n_2}-\theta_0|>$

$\delta_{n_1n_2}u_{\alpha/1}$。若$H$为：$\theta\leqslant\theta_0$，则否定域可取为 $U_{n_1n_2}>\theta_0+\delta_{n_1n_2}u_\alpha$。

如果 $\delta^2_{n_1n_2}$ 依赖于总体分布（即使在H成立时），则我们要使用（3·23）。根据（3·22），为估计 $\sigma^2_{n_1n_2}$，需要估计 σ^2_{10} 和σ^2_{01}。这一点等到本节末尾再谈。顺便说一句：在一些重要例子中前一情况为多，因为 $\delta^2_{n_1n_2}$"分布无关"，常是典型的非参数方法的特点。

2. 区间估计　这时不论从（3·23）或（3·24）出发都可以。但如用前者，必须得到 σ^2_{10} 和 σ^2_{01} 的估计。如用后者，必须得到 $\mathrm{Var}(U_{n_1n_2})$的估计，且这种估计必须对一切F、G去做，而不止限于$F=G$时。所以，在假设检验的情况，如当$F=G$（或其他原假设）时 $\mathrm{Var}(U_{n_1n_2})$为已知常数，则估计 $\mathrm{Var}(U_{n_1n_2})$ 的任务可免除，但在区间估计情况，纵使这一点成立，仍不能免除估计方差的问题。

例 3·7　再考虑例 3·3。根据（3·24）及(3·16)，知当原假设$F=G$成立，且当 $n_1\to\infty$，$n_2\to\infty$时，有

$$\sqrt{12n_1n_2}\left(U_{n_1n_2}-\frac{1}{2}\right)\Big/\sqrt{n_1+n_2+1}\xrightarrow{\mathscr{L}}N(0,1),$$

于是得到检验问题

$H:F=G\longleftrightarrow K:G(x)\leqslant F(x)$对一切 x，且$F\neq G$的水平 α 的大样本否定域为

$$U_{n_1n_2}>\frac{1}{2}+\sqrt{\frac{n_1+n_2+1}{12n_1n_2}}u_\alpha,\tag{3·25}$$

这就是大样本单侧 Wilcoxon 检验。当 n_1,n_2 都不大时，可根据 $U_{n_1n_2}$ 在原假设下的确切分布定出界限。不少统计表中载有有关的表，例如可参看中国科学院数学研究所概率统计室所编的《常用数理统计表》。

例 3·4 的情况与此类似——当原假设 $F=G$ 成立时，Var（$U_{n_1n_2}$）也是"分布无关"，其计算略繁但不难，我们把它留给读者作为练习，

例 3·8 设某种元件寿命X的分布F处处连续，$F(0)=0$。记 $\bar{F}(x)=1-F(x)$，有

P（元件寿命至少尚有 s | 元件在时刻 t 尚未失效）

$=P(X\geqslant s+t\mid X>t)=\bar{F}(s+t)/\bar{F}(t)$，

设所考察的时间不太长，而可假定在这段时间内元件无老化作用，则上述条件概率应与从起始时刻 $t=0$ 处计算者相同，即

$$\bar{F}(s+t)/\bar{F}(t)=\bar{F}(s),\quad s>0,\quad t>0 \qquad (3\cdot26)$$

所以，若我们要检验(在一段时间内)"元件无老化"的原假设H，则相应到分布上这假设可写成(3·26)的形式。其对立假设K是"元件有老化"。这意味着当元件用了一段时间 t 以后，它至少再能用 s 这么久的概率，不如从一开始用 s 这么久的概率。这导致对立假设的分布表述：

$$K:\bar{F}(s+t)/\bar{F}(t)<\bar{F}(s),\quad s>0,\quad t>0. \qquad (3\cdot27)$$

现随机抽出 n 个元件，测得其寿命为 $X_1,\cdots,X_n$，它们是 X 的简单样本。要据以检验$H\longleftrightarrow K$。

本问题的总体分布族是

$\mathscr{F}=\{F:F(0)=0,\ F$ 在$(-\infty,\ \infty)$处处连续$\}$，

根据H，K的意义，由下式定义的

$$\theta(F)=\int_0^\infty\int_0^\infty[\bar{F}(s)\bar{F}(t)-\bar{F}(s+t)]dF(s)dF(t)$$

是衡量原假设H与对立假设K之间的差距的一适当指标：当H成立时 $\theta(F)=0$，否则 $\theta(F)>0$。这样，原检验问题转化为 H：$\theta(F)=0\longleftrightarrow K:\theta(F)>0$。因为

$$\int_0^\infty\bar{F}(t)dF(t)=\bar{F}(t)F(t)\Big|_0^\infty+\int_0^\infty F(t)dF(t)$$

$$=0+\int_0^1 udu=1/2,$$

以及

$$\int_0^\infty\bar{F}(s+t)dF(s)dF(t)=P_F(X_1>X_2+X_3).$$

事实上，固定 $X_2=s$，$X_3=t$，而在此条件下求事件 $\{X_1>X_2+X_3\}$ 的条件概率，结果为 $\bar{F}(s+t)$。再利用公式 $P_F(X_1>X_2+X_3)=E_F(P_F(X_1>X_2+X_3|X_2,X_3))$ 即得。由上述结果可知：

$$\theta(F)=\frac{1}{4}-P_F(X_1>X_2+X_3). \qquad (3\cdot28)$$

记 $g(x_1,x_2,x_3)=\frac{1}{4}-I(x_1>x_2+x_3)$，则 $g(X_1,X_2,X_3)$ 为 $\theta(F)$ 之一无偏估计。将 g 对称化，得对称核

$$h(x_1,x_2,x_3)=\frac{1}{4}-\frac{1}{3}\{I(x_1>x_2+x_3)+I(x_2>x_3+x_1)+I(x_3>x_1+x_2)\}.$$

以此为核的、基于样本 $X_1,\cdots,X_n$ 的 U 统计量 U_n，是 $\theta(F)$ 的最小方差无偏估计。经过繁复但不困难的计算，得知对此核而言，在原假设 H 成立时有 $\sigma_1^2=5/(432\cdot9)$。又此处 $m=3$，因此 $m^2\sigma_1^2=5/432$。于是据（3·18）有

$$\sqrt{n}U_n \xrightarrow{\mathscr{L}} N(0,5/432)，当H成立时。$$

据此得出 $H\longleftrightarrow K$ 的水平 α 大样本检验，其否定域为

$$U_n>\sqrt{5/(432n)}\,u_\alpha.$$

如以前曾指出的，在检验问题中，θ 的选择有一定的灵活度，不必只有一种方法。拿本例来说，一个看来也很合理的选择是

$$\tilde{\theta}(F)=\int_0^\infty\int_0^\infty[\bar{F}(s)\bar{F}(t)-\bar{F}(s+t)]^2dF(s)dF(t), \qquad (3\cdot29)$$

可以找到 $\tilde{\theta}(F)$ 的一个无偏估计，只依赖 $X_1,\cdots,X_6$。我们把这个并不困难的问题留给读者作为练习。在这个无偏估计的基础上用 U 统计量法检验 $H\longleftrightarrow K$，其过程与使用 $\theta(F)$ 时完全类似。这两个检验何者为优？这问题就不那么容易回答。一个考虑是比较二者的极限分布的方差，小者为优。但是，方差小可能是由于所选指标的“灵敏度”不高。可靠的比较要依据各检验在对立假设下的功效，可参看§4·2。

读者一定注意到：在原假设成立时，F 就是指数分布（有密度 $\lambda e^{-\lambda x} I(x>0)$，$\lambda>0$ 为参数）。因此，本问题无非就是一个检验一组样本是否抽自某一指数分布的问题，它也可以用通常的 χ^2 拟合优度检验法去处理。另外，本问题也可以用带参数的 Колмогоров 检验去做，见§4·6。

例 3·9 Kendall 的 τ 检验。

设 (X,Y) 为二维随机向量，(X_i,Y_i)，$i=1,\cdots,n$，为其简单样本。我们要检验原假设

$$H: X, \ Y \text{独立}。$$

在初等统计中，往往假定 (X, Y) 有二维正态分布 $N(a,b,\sigma_1^2,\sigma_2^2,\rho)$。这时，$X$，$Y$ 是否独立，取决于 ρ 是否为 0，而检验问题成为 $\rho=0\longleftrightarrow\rho\neq0$。如果事先根据问题性质知道，当不独立时必为正相关，则检验问题成为 $\rho=0\longleftrightarrow\rho>0$。此处我们并未对 (X, Y) 的分布形式作任何假定，故必须找一个能反映独立和不独立之间的差异的指标 $\theta=\theta(F)$（F 是 (X,Y) 的分布）. 假定对立假设为正相关，则一个看来合理的指标是

$$\theta=P_F((X_1-X_2)(Y_1-Y_2)>0) \qquad (3·30)$$

理由是这样的：当 X,Y 独立时，θ 显然为 $\frac{1}{2}$（假定总体分布 F 处处连续），这一点请读者自证。当 X,Y 为正相关时，X 的增加（下降）倾向于使 Y 也增加（下降）。因此若 $X_1>X_2(X_1<X_2)$，倾向于有 $Y_1>Y_2(Y_1<Y_2)$。这样，$(X_1-X_2)(Y_1-Y_2)$ 更倾向于大于 0，而 θ 之值将大于1/2。这样，我们的检验问题成为

$$H: X,Y \text{ 独立（这时 } \theta=\frac{1}{2}) \longleftrightarrow K: \theta>1/2。 \qquad (3·31)$$

读者应当注意本例与例 3·8 的差别。在例 3·8 中，当原假设不成立时按老化的定义应有 $\bar{F}(s+t)<\bar{F}(s)\bar{F}(t)$，于是在该处 $\theta>0$ 是一顺理成章的结果。唯有这样，检验问题｛无老化$\longleftrightarrow$有老化｝才可以说转化成$\{\theta=0\longleftrightarrow\theta>0\}$. 此处则不然："正相关"并

无确切含义。因此，我们在前面所作的推理也并未严格证明：正相关必导致（3·30）定义的 θ 必大于 $\frac{1}{2}$。故（3·31）真正的含义是：我们把“正相关”解释为 $\theta>\frac{1}{2}$——正相关容许很多解释，此是其中之一，其合理性按前述直观分析说得通。当对正相关作这种解释时，检验问题{独立⟷正相关}确实转化为(3·31)。另外，X，Y独立并不与 $\theta=\frac{1}{2}$ 等价，故H不能写为 $\theta=\frac{1}{2}$。

θ 的形式直接提供了其一个无偏估计：

$$h((X_1,Y_1),(X_2,Y_2))=I((X_1-X_2)(Y_1-Y_2)>0),$$

以之为核作成U统计量 U_n。若以 F_1 和 F_2 分别记X和Y的分布（F的边缘分布）。由F连续知 F_1,F_2 连续，因而有

$$\begin{aligned} h_1(x_1,y_1)&=P_F((x_1-X_2)(y_1-Y_2)>0)\\ &=F_1(x_1)F_2(y_1)+(1-F_1(x_1))(1-F_2(y_1))\\ &=1-F_1(x_1)-F_2(y_1)+2F_1(x_1)F_2(y_1). \end{aligned}$$

因为 F_1,F_2 连续，据定理 2·1，$F_1(X_1)$和 $F_2(Y_1)$都服从(0,1)均匀分布 $R(0,\ 1)$。在H成立时，$F_1(X_1)$和 $F_2(Y_1)$独立。利用这些事实，不难算得当原假设H成立时有

$$\sigma_1^2=\mathrm{Var}(h_1(X_1,Y_1))=1/36.$$

又此处$m=2$。于是据（3·18），在H成立之下有

$$\sqrt{n}(U_n-1/2)\xrightarrow{\mathscr{L}}N\left(0,\frac{1}{9}\right).$$

据此得出(3·31)的一个大样本检验为：当 $U_n>\frac{1}{2}+u_\alpha/(3\sqrt{n})$ 时否定原假设H，不然就接受H。

这个检验是 M.G.Kendall 在 1938 年引进的. Kendall 所用的指标不是（3·30）定义的 θ，而是 $\tau=2\theta-1$。这使在X，Y独立时有 $\tau=0$，对立假设为 $\tau>0$。如用 τ，相应的U统计量显然应为

$$\tilde{U}_n=2U_n-1. \tag{3·32}$$

与此相应，得出的大样本检验为：

{当$\tilde{U}_n>2u_\alpha/(3\sqrt{n})$时否定$H$，不然接受$H$}。

（3·32）在统计著作中有时称为 Kendall 的 τ 统计量。当n不大时，可以根据 τ 统计量在原假设下的确切分布去决定检验的临界值。为此制作有表，例如，可参看 D.J.Best 的«Extended Tables for Kendall's tau»（Biometrika 60，1973，p.429—30）。

如果对立假设并无方向性，则 Kendall 检验不合用。但U统计量的理论仍可用来处理这个问题。例如，衡量X，Y是否独立的一个显然合理的指标是

$$\bar{\theta}=\iint_{-\infty}^{\infty}[F(x,y)-F_1(x)F_2(y)]^2dF(x,y).$$

X，Y “独立” 及 “不独立” 分别等价于 “$\bar{\theta}=0$” 及 “$\bar{\theta}>0$”。且X,Y “愈不独立”，则$F(x,y)$与$F_1(x)F_2(y)$的差异愈是倾向于大，因之$\bar{\theta}$这个指标很合理。使用它，检验问题{X，Y独立$\longleftrightarrow X,Y$不独立}转化为

$$\bar{\theta}=0\longleftrightarrow\bar{\theta}>0,$$

要检验（3·33），关键在于找到$\bar{\theta}$之一无偏估计。这并不难，留给读者作为练习。

以上讨论的各例都是在原假设下 $\mathrm{Var}(U_n)$ 或 σ_1^2 为 “分布无关” 的。在有些场合这一点不成立。或者在作 θ 的区间估计时，需要得到 $\mathrm{Var}(U_n)$或 σ_1^2 的估计，该估计且须不仅当原假设成立时适用。这也不难用U统计量法作到。例如，按定义有

$$\sigma_1^2=\mathrm{Var}(h_1(X_1))=E(h_1^2(X_1))-\theta^2$$
$$=E\{h(X_1,X_2,\cdots,X_m)h(X_1,X_{m+1},\cdots X_{2m-1})\}-\theta^2$$

θ^2 可以用 U_n^2 去估计。至于第一项，由其表达式可知，以 $\bar{h}(x_1,x_2,\cdots,x_{2m-1})=h(x_1,x_2,\cdots,x_m)h(x_1,x_{m+1},\cdots,x_{2m-1})$ 为核（注意这不是对称核，即使 h 为对称）的U统计量（暂记为 V_n）是其一无偏估计。由此得出 σ_1^2 的一个估计为 $V_n-U_n^2$，这不仅在原假设成立时适用。应当注意，U_n^2 并非 θ^2 的无偏估计，故 $V_n-U_n^2$也

不是 σ^2 的无偏估计，如要得到无偏估计，可取 $h'(x_1,\cdots,x_{2m})=h(x_1,\cdots,x_m)\ h(x_{m+1},\cdots,x_{2m})$ 为核作成U统计量 W_n，则 W_n 为 θ^2 之无偏估计，因而 V_n-W_n 为 σ_1^2 之无偏估计。据U统计量的理论，在适当条件下，这就是 σ_1^2 的最小方差无偏估计。

有时，在原假设下 θ 之值固定为 θ_0，θ_0 已知，与原假设下总体分布无关，但 σ_1^2 并非分布无关。这时，可以用 $V_n-\theta_0^2$ 估划 σ_1^2。这对于假设检验的目的已够了，当然不适用于区间估计问题。

两样本U统计量的情况与此类似，细节留给读者自已去完成。

习　　题

3-1 (*a*) 以$\mathscr{F}$记一切期望存在有限的一维对称分布族。对每个 $F\in\mathscr{F}$，以 $\theta(F)$ 记其对称中心。设 $X_1,\cdots,X_n$ 为从F中抽出的简单样本。证明：若 $n\geqslant 3$，则 $\theta(F)$ 的最小方差无偏估计不存在。

(*b*) 可用U统计量法来处理 $\theta(F)$ 的估计问题。取核函数 $h(x_1,x_2,x_3)=\mathrm{med}(x_1,x_2,x_3)$，则 $h(X_1,X_2,X_3)$为 $\theta(F)$ 的无偏估计（第二章习题14）。于是，以 h 为核的U统计量似乎是 $\theta(F)$ 的最小方差无偏估计，而这与 (*a*)矛盾。试找出问题在那里。

3-2 证明方差的级是 2 。

3-3 (*a*) 设F为一维分布函数，则$\iint_{-\infty}^{\infty}F(x)\,dF(x)\geqslant\frac{1}{2}$（注意：$F(x)$右连续）。等号当且仅当$F$处处连续。对这个事实作一概率上的解释。又如 $F(x)$取为左连续的（即用 $P(X<x)$ 定义$F(x)$），情况如何。

(*b*) 设 $F(x,y)$ 为二维分布函数，处处连续。则$\iint_{-\infty}^{\infty}F(x,y)\,dF(x,y)$的值介于 0 和 $\frac{1}{2}$之间，实际上，可取 $\left[0,\ \frac{1}{2}\right]$ 内任

何值（这与一维情况大不相同）。对高于二维的情况此结论也成立。

(*c*) 设X，Y独立，各有分布函数$F(x)$和$F(x-\theta)$，而$\theta>0$。证明$P(Y>X)>\frac{1}{2}$。

3-4 以$\mathscr{F}$记一切其3阶矩有限的一维分布族。对$F\in\mathscr{F}$，以$\alpha_r(F)$记F的r阶原点矩。求$\alpha_2(F)\alpha_3(F)$的最小方差无偏估计。

3-5 在例3·8中，代替在那里引进的$\theta(F)$，也可以用

$$\tilde{\theta}(F)=\int_0^{\infty}\int_0^{\infty}[\bar{F}(s)\bar{F}(t)-\bar{F}(s+t)]^2dF(s)dF(t)$$

作为衡量与原假设差距的指标。试求出$\tilde{\theta}(F)$的最小方差无偏估计（假定样本大小足够大）

3-6 证明在X，Y独立且分布连续时，（3·30）式的θ为$\frac{1}{2}$。又通过例子证明：θ可取$[0,1]$上任何值。

3-7 可以找一个"全方位的"检验独立性的指标：

$$\tilde{\theta}(F)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[F(x,y)-F_1(x)F_2(y)]^2dF(x,y)$$

这里$F(x,y)$是(X,Y)的分布函数，而$F_1(x)$和$F_2(y)$分别是X和Y的（边缘）分布。泛函$\tilde{\theta}(F)$定义在由一切二维分布构成的集上。试找出$\tilde{\theta}(F)$的最小方差无偏估（假定样本大小足够大），并以它为根据作独立性的大样本检验。

3-8 证明例3·8中的$\sigma_1^2=5/3888$，并求出方差。

3-9 证明例3·4中的U统计量其实只与样本的秩有关。就是说，若知道了$Y_1,\cdots,Y_{n2}$在合样本$X_1,\cdots,X_{n1},Y_1,\cdots,Y_{n2}$中的秩$R_1,\cdots,R_{n2}$，则能定下$U$统计量之值。

3-10 举一个例子证明：若在两样本U统计量$U_{n_1n_2}$中，n_1和n_2只有一个趋于无穷，则$U_{n_1n_2}$不一定有相合性。

3-11 证明（3·20）式。

第四章　使用样本的秩的统计方法

关于样本的"秩"，以及使用秩而构造的统计方法，常称为"秩方法"，在前章中已多次提起过了，且讨论过一个以U统计量形式出现的例子——Wilcoxon 秩和检验．本章的目的是对这种方法给一个比较系统的讨论．

§4·1　基本性质与渐近分布

本节的目的是讨论秩统计量的一些初步概率性质，及线性秩统计量的基本极限定理，它们是把秩用于统计推断的方法和理论基础．

一、定义及基本性质

定义 4·1 设 $X_1,\cdots,X_n$ 为样本（不必独立或同分布），其值两两不同，称$R_i=\sum_{j=1}^{n}I(X_j\leqslant X_i)$为$X_i$在样本$X_1,\cdots,X_n$中的秩，$i=1,\cdots,n$．换句话说，若$X_{(1)}<\cdots<X_{(n)}$为$X_1,\cdots,X_n$的次序统计量，而$X_iX_{(R_i)}$，则$R_i$为$X_i$之秩．记$R=(R_1,\cdots,R_n)$．$R$，或其一部分分量，称为样本$X_1,\cdots,X_n$的秩统计量．更进一步，$R$的任何已知函数，例如$\sum_{i=1}^{n}i\log R_i$，$\sum_{i=1}^{n}i^2R_i$等，也称为秩统计量．换言之，秩统计量就是完全由样本的秩所决定的统计量．使用秩统计量的统计方法统称秩统计方法，或简称为秩方法．

特别重要的一种类型是线性秩统计量，它是形如$\sum_{i=1}^{n}c_ia(R_i)$的统计量，为$c_1,\cdots,c_n$为已知常数．使用线性秩统计量的方法称为线性秩方法，它构成目前常用的秩方法的主体．

前已指出，秩方法在非参数统计中占有极其重要的地位．与

此相应，本章内容在本教程中也就占有重要的地位，其原因可列举很多。例如，秩方法使用灵活，易于在各种检验问题中，从直观出发构造出检验统计量（秩方法主要用于检验问题）；线性秩统计量有完备的大样本理论；其在原假设下往往为分布无关；秩方法的使用，相对于其他方法而言，计算上不算很复杂；秩方法与常用的一些方法（ t 检验之类）相比，其性能不差等等. 这最后一点在本章中将作更具体的解释。

秩方法的历史，较近代的一般认为始自 1904 年C.Spearman关于秩相关的论文。1936 年，著名统计学家Hotelling 以及 Pabst发表了一个基于秩的检验独立性的方法。总观之，尽管在 1900-1945年期间是现代数理统计学从奠基到成熟的时代，出现了Fisher、Pearson、Neyman 等大师，但秩方法以至整个非参数统计进展不大. 1945 年 F.Wilcoxon 发表了其重要的秩和检验。它不仅在应用上有较大意义，且构成往后秩方法发展的动力和出发点。因此这项工作在秩方法发展史上可算是一个里程碑。

秩方法的发展依赖于其极限理论。第一个比较普遍的结果（同分布情况）属于 Wald 和 Wolfowitz（1944年），到1949年 Noether作了重要改进。1958年 Chernoff 和 Savage 首先对两样本情况作出一般结果，而六十年代 Hajek 的重要工作又作了大的推进，经过这些大家的工作，秩统计量的极限理论（主要是线性情况）达到相当完善的地步，足以应付统计的需要。六、七十年代以来以至如今，仍有一些学者从事这方面的研究工作，结果往精深方向发展，但忽视了其统计意义。

在这样的背景下，自五十年代初期以来，秩统计方法经历过一个较快的发展时期. 发展了一大批用于一、二样本及多样本问题，用于方差分析、回归分析、独立性和随机性检验等等的秩方法，也出现了若干专著。

在本节一、二、三段中，我们总假定样本 $X_1,\cdots,X_n$ 为独立同分布，其公共分布 F 处处连续。后面这个条件保证了：以概率1，

$X_1, \cdots, X_n$ 互不相同，因而秩的意义确定，且$R_1, \cdots, R_n$取 1 到 n 之值 1 次且仅 1 次．更进一步有以下的基本事实：

定理 4·1 以$(r_1, \cdots, r_n)$记$(1, \cdots, n)$的任一置换，这样的置换共有 $n!$ 个，则

$$P\Big((R_1, \cdots, R_n) = (r_1, \cdots, r_n)\Big) = 1/n!.$$

证明　由$X_1, \cdots, X_n$为 iid.，从对称的角度立即看出，形式地可如下论证：找 i_k，使 $r_{i_k} = k$，$k = 1, \cdots n$，则 $(i_1, \cdots i_n)$ 为 $(1, \cdots, n)$之一置换，故$(X_{i_1}, \cdots X_{i_n})$与$(X_1, \cdots, X_n)$同分布，以 R_j' 记 X_{ij} 在 $X_{i_1}, \cdots, X_{i_n}$ 中的秩，则$(R_1', \cdots, R_n')$应与$(R_1, \cdots, R_n)$同分布．故

$$P\Big((R_1, \cdots, R_n) = (r_1, \cdots, r_n)\Big) = P\Big((R_1', \cdots, R_n') = (1, \cdots, n)\Big)$$

$$= P\Big((R_1, \cdots, R_n) = (1, \cdots, n)\Big)$$

最后一个概率与$(r_1, \cdots, r_n)$无关，即所有 $n!$ 个这样的概率都取同一个值．这个值必为 $1/n!$．证毕．

这个定理指出，在 iid.与分布连续之下，秩的分布与总体分布无关．这是它在非参数统计中有用的根本原因，从这个基本事实出发，原则上就不难得到任何秩统计量的分布，例如

$$P(R_i = j) = 1/n, \quad j = 1, \cdots, n, \quad i = 1, \cdots, n \tag{4·1}$$

$$P(R_i = u, R_j = v) = \frac{1}{n(n-1)}, u, v = 1, \cdots, n, \ u \neq v, \quad i, j = 1, \cdots n, \ i \neq j \tag{4·2}$$

一般，对任一线性秩统计量 $L = \sum_{i=1}^{n} c_i a(R_i)$，有$P(L = a = d_a/n!$，其中 d_a 表示 $n!$ 个数$\sum_{i=1}^{n} c_i a(r_i)$ $\Big((r_1, \cdots, r_n)$遍取$(1, \cdots, n)$的一切置换$\Big)$中等于 a 的个数．可是，当 n 较大时，这种分布在形式上很繁又无规则，并不便于应用，因而考虑用其极限分布取代之．

利用（4·1）和（4·2），不难得到

$$E(L)=n\bar{c}\bar{a}, \tag{4·3}$$

$$\mathrm{Var}(L)=\frac{1}{n-1}\sum_{i=1}^{n}(a(i)-\bar{a})^2\sum_{i=1}^{n}(c_i-\bar{c})^2 \tag{4·4}$$

此处 $L=\sum_{i=1}^{n}c_i a(R_i)$，$\bar{a}=\sum_{i=1}^{n}a(i)/n$，$\bar{c}=\sum_{i=1}^{n}c_i/n$。

（4·3）不难证明，留给读者自证。为证（4.4），只须注意

$$\mathrm{Var}\left(a(R_i)=\sum_{i=1}^{n}\right)(a(i)-\bar{a})^2/n,\quad i=1,\cdots,n$$

及

$$\mathrm{Cov}\left(a(R_i),a(R_j)\right)=\frac{1}{n(n-1)}\sum_{u\neq v}^{n}a(u)a(v)-\bar{a}^2$$

$$=\frac{-1}{n(n-1)}\sum_{i=1}^{n}\left(a(i)-\bar{a}\right)^2,\quad i\neq j,$$

用公式

$$\mathrm{Var}\,(L)=\sum_{i=1}^{n}c_i^2\mathrm{Var}\left(a(R_i)\right)+\sum_{i\neq j}^{n}c_ic_j\mathrm{Cov}\left(a(R_i),a(R_j)\right)$$

稍加整理即得。

例 4·1 考察例 3·4 中提到过的 Wilcoxon 两样本秩和统计量在 $F=G$ 的情形，且 F 处处连续。即：$X_1,\cdots,X_{n1}$，$Y_1,\cdots,Y_{n2}$ 为独立同分布，以 R_{n_1+i} 记 Y_i 在合样本 $X_1,\cdots,Y_{n2}$ 中的秩，$L=\sum_{i=1}^{n2}R_{n1+i}$ 这相当于

$$a(i)=i,\quad i=1,\cdots,n_1+n_2;\quad c_i=0 \text{ 当 } i\leqslant n_1,\quad c_i=1 \text{ 当 } n_1+1\leqslant i\leqslant n_1+n_2 \tag{4·5}$$

记 $N=n_1+n_2$。易算出

$$\sum_{i=1}^{N}\left(a(i)-\bar{a}\right)^2=N(N^2-1)/12,\ \sum_{i=1}^{N}(c_i-\bar{c})^2=n_1n_2/N,$$

于是得到

$$E(L)=\frac{n_2(N+1)}{2},\quad \mathrm{Var}(L)=\frac{n_1n_2(N+1)}{12}. \tag{4·6}$$

这与（3·16）一致（注意此处的 L 是该处的 $U_{n_1n_2}$ 的 n_1n_2 倍再加

上 $n_2(n_2+1)/2$。

下面两个定理涉及线性秩统计量分布的对称性。

定理 4·2 对线性秩统计量 $L=\sum_{i=1}^{n}c_i a(R_i)$，若以下两条件至少成立其一：

$$a(i)+a(n+1-i)=a(1)+a(n),\ i=1,\cdots,n, \tag{4·7}$$

$$c_i+c_{n+1-i}=c_1+c_n,\quad i=1,\cdots,n, \tag{4·8}$$

则 L 的分布关于其期望 $n\bar{a}\bar{c}$ 对称。

证明 据定理 4·1 有

$$(n+1-R_1,\cdots,n+1-R_n)\overset{d}{=}(R_1,\cdots,R_n), \tag{4·9}$$

若（4·7）成立，则$a(R_i)-\bar{a}=\bar{a}-a(n+1-R_i)$.故由（4·9），

$$L-n\bar{a}\bar{c}=\sum_{i=1}^{n}c_i\Big(a(R_i)-\bar{a}\Big)=\sum_{i=1}^{n}c_i\Big(\bar{a}-a(n+1-R_i)\Big)$$

$$\overset{d}{=}\sum_{i=1}^{n}c_i\Big(\bar{a}-a(R_i)\Big)=n\bar{a}\bar{c}-L,$$

即 $L-n\bar{a}\bar{c}$ 与 $-(L-n\bar{a}\bar{c})$同分布，因而 $L-n\bar{a}\bar{c}$ 的分布关于 0 对称，这证明了所要的结果。当（4·8）成立时证明类似，留给读者。

若分布对称，则在造表时可以只考虑一端。故本定理在应用上有意义。下面是一个更广一些的结果。

定理 4·3 以R记$(R_1,\cdots,R_n)$，$\mathscr{F}$记（1，…，n）的所有置换之集（$\mathscr{F}$包含 $n!$ 个元素）。设 f 为由$\mathscr{F}$到$\mathscr{F}$上的一个一一对应变换。设V为定义在$\mathscr{F}$上的实函数，满足条件

$$V(r)+V\Big(f(r)\Big)=\text{常数}\ c，\text{对一切}\ r\in\mathscr{F}, \tag{4·10}$$

则统计量 $V(R)$ 的分布关于 $c/2$ 对称。此断言之逆亦真。

证明 因f为由$\mathscr{F}$到$\mathscr{F}$上的一一对应变换，由定理 4·1 知，$f(R)\overset{d}{=}R$.故若（4·10）成立，则有

$$V(R)-c/2=c/2-V(f(R))\overset{d}{=}c/2-V(R)$$

因而得出 $V(R)-c/2$ 关于0对称。反过来，若 $V(R)$ 之分布关于 $c/2$ 对称，任取 $a>c/2$，使 $P(V(R)=a)>0$。记 $d=a-c/2$。则由 $V(R)$ 的分布关于 $c/2$ 对称可知

$$P\Big(V(R)=c/2-d\Big)=P\Big(V(R)=c/2+d\Big)>0.$$

由此，再注意到 $V(R)$ 以等概率取 $\mathscr{F}$ 上每一元为值，知两集合 $\{r:r\in\mathscr{F};\ V(r)=c/2+d\}$ 和 $\{r:r\in\mathscr{F};\ V(r)=c/2-d\}$ 所含元素个数相同，故在这两集之间可建立一一对应。因对不同的 $a>0$ 集合 $\{r:r\in\mathscr{F};\ V(r)=a\}$ 互不相交，对一切 $a>0$ 建立上述对应，从而在整个 $\mathscr{F}$ 上建立了一一对应 f。这个对应显然满足（4·10）。定理证毕。

不难验证：定理4·2是本定理的特例，细节留给读者。

二、同分布下线性秩统计量的渐近正态性

本段仍然假设样本 $X_1, \cdots, X_n$ 是独立同分布，而且其公共分布 F 连续，$R=(R_1,\cdots,R_n)$ 为秩统计量。考虑线性秩统计量 $L_n=\sum_{i=1}^{n}c_{ni}a_n(R_i)$。这里因为要考虑样本大小 $n\to\infty$ 时的情况，我们把前面用过的记号 $c_i, a(R_i)$ 和 L 都添上足标 n。在这个表达式中，$a_n(\cdot)$ 为一个定义在集合 $\{1,2,\cdots,n\}$ 上的实函数，有时称它为计分函数。道理是这样的：$a_n(R_i)$ 愈大，这一项在 L_n 中起的作用也愈大，形象地可以说成是 R_i "得了 $a_n(R_i)$ 分"。$c_{n1},\cdots,c_{nn}$ 为常数，它们有时被称为"回归系数"。这个名词在很大程度上是借用性质的。

记 $l_n=E(L_n)$，$\sigma_n^2=\mathrm{Var}(L_n)$。问题是要探究，当 $\{c_{ni}\}$ 及函数 a_n 满足何种条件时，标准化后的变量 $(L_n-l_n)/\sigma_n$ 依分布收敛于 $N(0,1)$。这个问题曾费了不少知名的统计学家的心血，而以 Hajek 1961 年的工作最完整(见他发表在 Ann.Math.Statist.上的文章，1961年 p.506)。他的记法很富技巧性且基本上只用了

初等工具。因为太繁，这里只好从略。而且，就用于统计推断而言，最一般形式的 Hajek 定理并不方便，倒是由之推出的两个结果，适用于许多问题.故我们这里只限于不加证明地引述这两个结果.作为准备，要引进几个概念。

设对每个自然数 n 给定了 n 个实数 $c_{n1},\cdots,c_{nn}$。记 $\bar{c}_n=(c_{n1}+\cdots+c_{nn})/n$。

定义 4·2 如果当 $n\to\infty$时，有

$$\max_{1\le i\le n}(c_{ni}-\bar{c}_n)^2\Big/\sum_{i=1}^{n}(c_{ni}-\bar{c}_n)^2\longrightarrow 0, \qquad (4\cdot 11)$$

则称序列$\{(c_{n_1},\cdots,c_{nn}):n=1,2,\cdots\}$满足条件$N$(条件$N$之得名，是因为此条件是 Noether 在 Wald-Wolfowitz 1944 年一个类似条件的基础上，于1949年引进的).有时，我们只是对自然数的一个子序列$\{n_1,n_2,\cdots\}$中的 n 给定了$\{c_{n_1},\cdots,c_{nn}\}$。这时定义(4·2)仍有效，但把$n\to\infty$改为 n_i 中的 $i\to\infty$。

细察(4·11)看出：这个条件无非是说，在当 $n\to\infty$时，构成平方和$\sum_{i=1}^{n}(c_{ni}-\bar{c}_n)^2$ 的每一项所起的作用，一致地趋向于0。这与在中心极限定理中起关键作用的所谓“一致渐近可忽略”的条件是一种性质。由此也就可以理解：为什么这样的条件N会出现在 L_n 的渐近正态性的讨论中。

其次引进一个函数类 SS，它由一切定义于(0,1)区间上的满足下述条件的函数 φ 构成：$\varphi=\varphi_1-\varphi_2$，$\varphi_1$.$\varphi_2$ 都是定义在(0,1)的，非降而平方可积的函数，且 φ_1，φ_2 在(0,1)区间内都不恒等于常数。

定理 4·4[①] 对线性秩统计量$L_n=\sum_{i=1}^{n}c_{ni}a_n(R_i)$，若下述两条

① 参看§5·2一、及第五章附录，其中将给出本定理重要特例的证明。

件满足：

(1) $\{(c_{n1},\cdots,c_{nn}):n=1,2,\cdots\}$满足条件$N$.

(2) 存在常数 $h_n\neq 0$ 及函数 $\varphi\in SS$，使

$$a_n(i)=h_n\varphi\left(\frac{i}{n+1}\right),\quad i=1,\cdots,n \tag{4·12}$$

则当 $n\to\infty$时，有

$$(L_n-l_n)/\sigma_n \xrightarrow{\mathscr{L}} N(0,1). \tag{4·13}$$

此处 $l_n=E(L_n)$，$\sigma_n^2=\mathrm{Var}(L_n)$，已在前面提到过。

说来有趣，本定理（以至 Hajek 的一般定理）的证法也是基于在前几章中多次提到的那个原则，即要设法把 L_n 表为一个独立随机变量和加上一个余项，后者的影响当 $n\to\infty$时趋于 0。然而，这原则说起来简单，实际做起来却大有文章。本定理的证明是很好的例子。

例 4·2 再考察 Wilcoxon 秩和统计量 L_n。严格说来，Wilcoxon统计量依赖于两个分样本大小 n_1,n_2，应记为$L_{n_1n_2}$才确切，此处我们以其合样本大小 $n=n_1+n_2$ 为足标。只要注意到这一点，当不致引起混淆，此处有

$$(c_{n1},\cdots,c_{nn})=(0,\cdots0,1,\cdots,1)(n_1\text{ 个 }0,n_2\text{ 个 }1) \tag{4·14}$$

不难算出$\sum\limits_{i=1}^{n}(c_{ni}-\bar{c}_n)^2=n_1n_2/n$，而 $\max\limits_{1\le i\le n}(c_{ni}-\bar{c}_n)^2<1$。因此，只要 $n_1n_2/n\to\infty$，则条件 N 满足。不难看出，此条件等价于

$$n_1\to\infty,\quad n_2\to\infty. \tag{4·15}$$

因此，在两个分样本大小都无限增加时，条件N满足。在本例中，n 不一定跑遍全部自然数，这在定义 4·2 后面已有所交代。

其次，若令 $\varphi(u)=u$ $(0<u<1)$，而 $h_n=n+1$，则 $a_n(i)=i=h_n\varphi\left(\frac{i}{n+1}\right)$，$i=1,\cdots,n$。$\varphi$ 可以表为$2u-u$，其中 $2u$ 和u 都是在（0,1）非降非常数的平方可积函数，于是 $\varphi\in SS$、再用

例 4·1 中算得的 $E(L_n)$ 和 $\mathrm{Var}(L_n)$，得知当 $n_1\to\infty$，$n_2\to\infty$ 时，有

$$2\sqrt{3}\left(L_n-\frac{n_2(n+1)}{2}\right)/\sqrt{n_1n_2(n+1)}\xrightarrow{\mathscr{L}}N(0,1).\quad(4\cdot16)$$

这个结果就可用于大样本检验，这在例 3·7 中已讨论过了．在该例中曾用 U 统计量的理论求得 L_n 的极限分布，但多数情况下秩统计量并不一定能表为 U 统计量，故此例只能算是一个巧合．

本定理中的计分函数形式 $\left(a_N(i)=h_n\varphi\left(\frac{i}{n+1}\right)\right)$，是最重要、应用最广的一种形式．除此之外，还有一类计分函数也很重要，即在下一定理中所涉及的内容．

设有一个一维分布 D，而 $V_{n1}\leqslant\cdots\leqslant V_{nn}$ 是从此分布中抽出的，大小为 n 的次序样本．令

$$a_n(i)=E(V_{ni}),\ i=1,\cdots,n,\quad(4\cdot17)$$

以之作为计分函数(此处自然要求分布 D 的期望存在有限)．这种计分函数最早且最著名的一个，是 Fisher 和 Yates 在 1938 年提出的，他们取 D 为标准正态分布 $N(0,1)$，并对 $n\leqslant50$ 给出了在此分布下(4·17)右边之值（见 R.A.Fisher and F. Yates, Statistical Tables for Biological, Agricultural and Medical Research, Oliver and Boyd., 1938)，易见 Wilcoxon 统计量的计分函数也有这个形式，其中 D 为(0,1)均匀分布(差一个无关紧要的常数倍数)．

如果分布函数 D 在 $(-\infty,\infty)$ 处处严格增加，则 D 的反函数 D^{-1} 存在，而(4·17)可写成另外的形式：

$$a_N(i)=E(V_{ni})=E(D^{-1}(U_{ni})),\quad(4\cdot18)$$

此处 $U_{n1}\leqslant\cdots\leqslant U_{nn}$ 为抽自(0,1)均匀分布的次序样本（参看定理 2·1)．作为(4·17)的一个稍稍的推广，我们把(4·18)中的 D^{-1} 改为 φ，而不要求 φ 是某一分布函数的反函数．

定理 4·5 若 $L_n=\sum_{i=1}^{n} c_{ni}a_n(R_i)$，其中

$\{(c_{n1},\cdots,c_{nn}):n=1,2,\cdots\}$满足条件$N$，又 $a_N(i)=E\left(\varphi(U_{ni})\right)$，$1\leqslant i\leqslant n$，而$\varphi\in SS$。则当 $n\to\infty$时，（4·13）成立。

例 4·3 仍考察两样本问题，以 $X_1,\cdots,X_{n1}$ 和 $Y_1,\cdots,Y_{n2}$ 分别记从一维分布F和G中抽出的简单样本。设$F=G$，F处处连续，以 R_{n_1+i} 记Y_i 在合样本中的秩，$i=1,\cdots,n_2$。取计分函数为（4·17），其中分布D为$N(0,1)$，又（$c_{n1},\cdots,c_{nn}$）按（4·14）定义。由此得出的线性秩统计量就是 $L_n=\sum_{i=1}^{n_2}E\left(\Phi^{-1}(U_{n,R_{n1+i}})\right)$，$\Phi$ 为$N(0,1)$ 的分布。这样定义的 L_n，称为 Fisher-Yates 统计量。

当 $n_1\to\infty$，$n_2\to\infty$时，例 4·2 已证明：$\{(c_{n1},\cdots,c_{nn})$：$n=1,2,\cdots\}$ 满足条件 N。按定理 4·5，为证L_n 渐近正态，还必须证明

$$\int_0^1\left(\Phi^{-1}(u)\right)^2du<\infty. \tag{4·19}$$

为证此，以 φ 记$N(0,1)$的密度，经作变换$u=\Phi(x)$，知上式等价于$\int_{-\infty}^{\infty}x^2\varphi(x)dx<\infty$。但此积分就是$N(0,1)$的方差，即1，故为有限。因而证明了（4·19）。(此证明可推广为：若 a_N（·）由（4·17）定义，分布函数D处处严增，且D的方差有限，则$a_N(i)$满足定理 4·5 的条件。)

在讨论多样本问题及其他问题中，须用到几个线性秩统计量的联合分布收敛于多维正态。关于这个问题，我们只不加证明地引述下列定理。

定理 4·6 仍在样本独立同分布及公共分布处处连续的假定下，考察m个线性秩统计量：

$$L_{nk}=\sum_{i=1}^{n}c_{ni}^{(k)}a_n(R_i),\quad k=1,\cdots,m.$$

设以下条件成立：

(1) 对每个 k，$\{(c_{n1}^{(k)}\cdots,\ c_{nn}^{(k)}):n=1,2,\cdots,\cdots\}$ 满足条件N

(2) 计分函数 $a_n(\cdot)$满足定理 4·4 或者定理 4·5 的条件。

(3) 记 $\bar{c}_n^{(k)}=\sum_{i=1}^{n}c_{ni}^{(k)}/n$，$k=1,\cdots,m$，则对任何$k\neq l$，$k\leqslant m$，$l\leqslant m$，极限

$$\lim_{n\to\infty}\frac{\sum_{i=1}^{n}\left(c_{ni}^{(k)}-\bar{c}_n^{(k)}\right)\left(c_{ni}^{(l)}-\bar{c}_{ni}^{(l)}\right)}{\sqrt{\sum_{i=1}^{n}\left(c_{ni}^{(k)}-\bar{c}_n^{(k)}\right)^2\sum_{i=1}^{n}\left(c_{ni}^{(l)}-\bar{c}_n^{(l)}\right)^2}}=\lambda_{kl}$$

存在，且方阵 $\Lambda=(\lambda_{kl})_{k,l}=1,\cdots,m$ 为满秩(注意 $\lambda_{kk}=1$)。

记 $l_{nk}=E(L_{nk})$，$\sigma_{nk}^2=\mathrm{Var}(L_{nk})$。则当 $n\to\infty$时有

$$\left(\frac{L_{n1}-l_{n1}}{\sigma_{n1}},\ \cdots,\ \frac{L_{nm}-l_{nm}}{\sigma_{nm}}\right)'\xrightarrow{\mathscr{L}}N(0,\Lambda). \qquad (4\cdot20)$$

本定理的应用实例将在 §4·4 中讨论。

定理 4·4 和 4·5 指明，在一定的条件下，线性秩统计量 L_n 经标准化后，$(L_n-l_n)/\sigma_n$ 的分布函数收敛于标准正态分布函数。对固定的n这二者的差距如何？因为当 n 较小时，$(L_n-l_n)/\sigma_n$ 的分布，易根据定理 4.1 算出，可以拿计算结果与标准正态分布比较，而对上述差距得到一些概念。如果当 n 较小时这差距尚不大，则有理由相信，当 n 更大时，逼近的程度当更好。Lehmann 曾在其著作《Nonparametric Statistical Methods Based On Ranks》中引述了 Wilcoxon 两样本秩和统计量的结果：按定理 4·4，有

$$P(L_n\leqslant c)\approx\Phi\left(\frac{c-n_2(n+1)/2}{\sqrt{n_1n_2(n+1)/12}}\right) \qquad (4\cdot21)$$

$n=n_1+n_2$，Φ 为$N(0,1)$的分布。因 L_n 只取整数值，一般作连续性修正：

$$P(L_n \leqslant c) \approx \Phi\left(\frac{c - n_2(n+1)/2 + 1/2}{\sqrt{n_1 n_2 (n+1)/12}}\right) \qquad (4.22)$$

对几组（n_1，n_2）及c之值，下表给出$P(L_n \leqslant c)$的确值，以及（4·21）和（4·22）的右边的值：

$n_1 = 6$，　　$n_2 = 3$

c	6	7	8	9	10
确值	·012	·024	·048	·083	·131
（4·21）	·010	·019	·035	·061	·098
（4·22）	·014	·026	·047	·078	·123

$n_1 = 12$，　　$n_2 = 4$

c	13	15	20	23	25
确值	·004	·010	·052	·106	·158
（4·21）	·005	·011	·045	·091	·138
（4·22）	·006	·012	·051	·102	·151

$n_1 = 8$，　　$n_2 = 8$

c	44	46	48	52	56
确值	·005	·010	·019	·052	·117
（4·21）	·006	·010	·018	·047	·104
（4·22）	·007	·012	·020	·052	·114

从这几个表来看，用正态分布逼近 Wolcoxon 统计量的分布，即使对比较小的n_1, n_2，效果还是比较好，特别是（4·22），其误差从实用的观点看已无甚重要性。相信对常用的一些线性秩统计量，情况应基本相当。

三、样本独立同分布但有结存在时

如果样本$X_1, \cdots, X_n$中有相同的，则它们构成一个"结"(tie)，结中样本个数称为该结的长。例如，设有样本

0.45，0.20，0.80，0.20，0.34，0.45，0.15，0.56，0.20（4·23）

则其中有两个结：X_2, X_4, X_9都为0.20，此结之长为3；X_1和

X_6 都为0.45，此结之长为 2. 为方便计，有时也把不重复的样本称为长为 1 的结。这样，此样本中有 4 个长为 1 的结，即 X_3，X_5，X_7和 X_8。

当结（长大于 1 者）出现时，样本的秩如何定，需要另加明确。且一般说来，结的存在使秩方法的理论和实际使用都复杂化了。由于这个原因，人们往往假定总体分布连续，以回避这个问题。但有时问题的性质使连续性假定不合适，另外，即使总体分布连续，也可以由于测量单位较粗而出现结，例如，两个样本较精细之值本应为 3.1413 和 3.1426。但如只记到小数点后两位，则都为3.14，而形成结。故在研究秩方法时，对结的问题作出适当处理是有其必要的。

常见的处理结的做法有以下两种：

1. 随机化法。就是把同一个结内的样本，按该结所占位置，用机会均等的方法配给其秩。例如，样本（4·23）中的 X_2，X_4 和 X_9 构成一结，它们占据了 2 、3 、4 这三个位次（秩）。按"抽签"的方法，把这三个秩随机地分配给 X_2，X_4 及 X_9。同样，X_1 和 X_6 这个结占据位次 6 和 7，可投掷一均匀铜板，如出现正(反）面，则把秩 7（6）赋予 X_1，余下那一个给 X_6。

采用这种方法定秩，最大的优点在于定理 4·1 的结论仍成立：

定理 4·1′ 设 $X_1,\cdots,X_n$ 为从一维分布 F 中抽出的简单样本。不论 F 是否连续，若按上述方法决定 X_i 之秩为 R_i，$i=1,\cdots,n$，则秩统计量 $R=(R_1,\cdots,R_n)$ 的分布仍如定理 4·1 所示。

证 以 $U_1,\cdots,U_n$ 记 (0,1) 均匀分布的简单样本，且设 $X_1,\cdots,X_n$，$U_1,\cdots,U_n$ 全体独立。记 $Y_i=(X_i,U_i)$，$i=1,\cdots,n$。对 $Y_1,\cdots,Y_n$ 排序如下：任取 $i\neq j$。若 $X_i<X_j$，则 Y_i 在 Y_j 之前。若 $X_i=X_j$ 但 $U_i<U_j$，则 Y_i 在 Y_j 之前。由于以概率1，$U_1,\cdots,U_n$ 互不相同，上述规则唯一地决定了 $Y_1,\cdots,Y_n$ 的排序，因而唯一决定了 $Y_1,\cdots,Y_n$ 之秩，它们显然也就是 $X_1,\cdots,X_n$ 按上述随机化方法决定的秩 $R_1,\cdots,R_n$。但就 $Y_1,\cdots,Y_n$ 而言，定理

4·1的推理完全适用。因而证明了所要的结果。

读者应注意的是：当我们说$Y_1,\cdots,Y_n$之秩就是$X_1,\cdots,X_n$按随机化方法决定的秩时，所指的是，通过（0,1）均匀变量U_1，$\cdots,U_n$以在诸X_i相同时施行随机化，确符合“结中诸变量的秩按机会均等”的方式给秩的要求。你可以采用别的机制实现随机化，但所得秩统计量的分布都一样。

有了这个结果，前面一、二两段在总体分布连续的条件下得出的一切结果，在此全保持成立。例如，在原假设（两总体的分布F，G相同）下，两样本 Wilcoxon 秩和统计量的分布及其极限定理，与以前求得者相同，因而检验的临界值也一样。

这个方法虽然从理论的角度说颇简单，但有一个根本的缺陷，就是引进了一个外来的，人为的随机化手续。这就使得甲、乙两人在同一组样本之下，由于这随机化结果之不同，而得出不同的秩统计量值。举例而言，设在样本（4·21）中，X_1,X_3,X_8来自第二总体（照以前记法是Y样本），其余来自第一总体，而我们打算用 Wilcoxon 秩和检验去检验“两总体同分布”之假设H。以W记Y样本之秩和。设根据检验水平而确定的临界值是：$W\geqslant 24$时否定H，不然就接受H。现Y样本中有两个之秩分别为 8 和 9，另一个属于一长为 2 的结。因此其秩或为 6 或为 7，要看随机化的结果如何。若甲施行随机化的结果给予这个Y样本以秩 7，则$W=24$而甲否定H。同时，乙施行随机化给以秩 6，则$W=23$，而乙接受了H。对应用者来说这很难于接受。下面的“平均法”就没有这个缺点。

2. 平均法。此法对结中每一样本赋予均等之秩，即结中各位置秩的平均。拿样本（4·23）而言，X_2,X_4,X_9这个结占据位次 2，3，4，其秩平均为$(2+3+4)/3=3$。故R_2，R_4和R_9都定为 3。同样，X_1，X_6这个结占据位次 6 和 7，其平均为$(6+7)/2=6.5$。故R_1和R_6都定为 6.5。剩下的R_3,R_5,R_7，R_8分别为 9，5，1 和 8。这个做法达到了样本之秩唯一决定的

目的，取平均在直观上看也是自然的。至于其缺点，则在于这样决定的秩统计量，其分布已不适合定理4·1(这是显然的，因现在秩可以取非整数)。实际上，在平均法之下，即使对简单样本而言，如总体分布可以不连续，则秩统计量之分布将依赖于总体分布，即并非分布无关。这样，例如对两样本而言，已不可能根据检验水平去决定秩检验统计量的临界值，换句话说，在平均法之下，如允许总体分布不连续，则秩检验已不能认为是在典型意义下的非参数方法。但是，二、中的极限定理经过适当的修改后仍成立，这提供了大样本检验的可能性。

为叙述这种极限定理，引进所谓“结统计量”是有益的，简言之，结统计量记载了样本中各结之长(包括长为1之结)。拿样本(4·23)为例，按由小到大排列为

0.15，0.20，0.20，0.20，0.34，0.45，0.45，0.56，0.80，共有6个结，其长分别为 $\tau_1=1$，$\tau_2=3$，$\tau_3=1$，$\tau_4=2$，$\tau_5=1$，$\tau_6=1$。故在此例，结统计量为 $\tau=(\tau_1,\cdots,\tau_6)=(1,3,1,2,1,1)$。一般地，若样本 $X_1,\cdots,X_n$ 的排列是：

$$
\begin{aligned}
&X_{i_1}=X_{i_2}=\cdots=X_{i\tau_1}\\
&\quad<X_{i\tau_1+1}=X_{i\tau_1+2}=\cdots=X_{i\tau_1+\tau_2}\\
&\quad<\cdots\cdots\\
&\quad<X_{i\tau_1+\cdots+\tau_{q-1}+1}=\cdots=X_{i\tau_1+\cdots+\tau_q}\quad(\tau_1+\cdots+\tau_q=n),
\end{aligned}
$$

则结统计量为$\tau=(\tau_1,\cdots,\tau_q)$。注意 q 与样本有关，因而为随机的。又结统计量 τ 并未指明那些样本在那个结内。如对(4·21)而言，若有人(他知道这样本的具体值)告诉你这样本的结统计量为(1，3，1，2，1，1)，你无法据此知道在长为3的结中包含了那三个样本。

其次，要根据平均法的精神，对线性秩统计量的定义 $\sum_{i=1}^{n}c_ia(R_i)$ 加以修改。方法如下：设有样本 $X_1,\cdots,X_n$，按由小到大排列为(4·24)。按平均法，对样本(4·24)而言，秩只取

q 个值：

$$\begin{aligned} d_1 &= (1+2+\cdots+\tau_1)/\tau_1 = (1+\tau_1)/2, \\ d_2 &= (\tau_1+1+\cdots+\tau_1+\tau_2)/\tau_2 = \tau_1+(1+\tau_2)/2, \\ d_q &= (\tau_1+\cdots+\tau_{q-1}+1\cdots+\tau_1+\cdots+\tau_q)/\tau_q \\ &= \tau_1+\cdots+\tau_{q-1}+(1+\tau_q)/2, \end{aligned} \tag{4·25}$$

把函数$a(\cdot)$在每个结上取的值加以平均，也得 q 个值：

$$\begin{aligned} t_1 &= \Big(a(1)+\cdots+a(\tau_1)\Big)/\tau_1, \\ t_q &= \Big(a(\tau_1+\cdots+\tau_{q-1}+1)+\cdots+a(\tau_1+\cdots+\tau_q)\Big)/\tau_q. \end{aligned} \tag{4·26}$$

定义新函数$\tilde{a}(\cdot)$：

$$\tilde{a}(d_i)=t_i, \quad i=1,\cdots,q, \tag{4·27}$$

而将线性秩统计量的原定义 $L=\sum_{i=1}^{n}c_i a(R_i)$ 修改为：

$$\tilde{L}=\sum_{i=1}^{n}c_i\tilde{a}(\tilde{R}_i), \tag{4·28}$$

此处 $\tilde{R}_i$ 为 X_i（在样本 $X_1,\cdots,X_n$ 中的）按平均法决定的秩，（4·28）的定义过程似颇复杂，其实质很简单：把原来每个 $a(R_i)$ 按结上$a(\cdot)$的平均值取代之。

例 4·4　设有两组样本：X 样本和Y 样本，大小都为10。具体值为：

X 样本：7，6，7，5，4，6，5，6，6，5，

Y 样本：5，6，6，3，4，7，4，5，5，6，要计算 Wilcoxon 秩和统计量W，即Y样本在平均法之下的秩和。

先把合样本按由小到大排列，并以 * 标出Y 样本（* 号标在结中何处无关紧要）：

3*，　4*，4*，4，　5*，5*，5*，5，5，5，

$\tau_1=1$　$\tau_2=3$　$\tau_3=6$

6*，6*，6*，6，6，6，6，　7*，7，7，

$\tau_4=7$　$\tau_5=3$

$d_1=1$，$d_2=3$，$d_3=15/2$，$d_4=14$，$d_5=19$，

因$a(i)=i$，标出 $t_i=d_i$，$i=1,\cdots,5$，于是$\bar{a}(\cdot)$为$\bar{a}(d_i)=t_i$，$i=1,\cdots,5$。按平均法，10个Y样本所占之秩分别为1，3，3，15/2，15/2，15/2，14，14，14，19。W是它们的和，即90.5。形式地按（4·28）算，则须先定义

$c_1=\cdots=c_{10}=0$，$c_{11}=\cdots=c_{20}=1$

然后按去$\widetilde{R}_{11}=1$，$\widetilde{R}_{12}=3$，$\cdots$，$\widetilde{R}_{20}=19$去计算。

另外，定义以下两个量，它相当于由公式（4·3）和（4·4）给出的$E(L)$和$\mathrm{Var}(L)$：

$$m_n=n\bar{c}\bar{a},\ \sigma_n^2(\tau)=\frac{1}{n-1}\left(\sum_{i=1}^{q}\tau_i t_i^2-n\bar{a}^2\right)\sum_{i=1}^{n}(c_i-\bar{c})^2 \qquad (4\cdot29)$$

其中 $\bar{c}=\sum_{i=1}^{n}c_i/n$，$\bar{a}=\sum_{i=1}^{n}a(i)/n$。

现设有一串线性秩统计量（按平均法定义）：

$$\widetilde{L}_n=\sum_{i=1}^{n}c_{ni}\bar{a}_n(\widetilde{R}_i),\ n=1,2,\cdots,$$

对每个固定的n，定义（4·29）的 m_n 和 $\sigma_n^2(\tau)$。即把 c_i 改为 c_{ni}，$a(i)$改为 $a_n(i)$ 去计算 $\bar{c}_n$ 和 $\bar{a}_n$，以之取代（4·29）中的 $\bar{c}$ 和 $\bar{a}$。又这时 τ，q，t_i 等当然也与 n 有关。

定理4·4′ 设$\{(c_{n1},\cdots,c_{nn}):n=1,2,\cdots\}$及$\{a_n(\cdot):n=1,2,\cdots\}$满足定理 4·4 的条件，则当 $X_1,\cdots,X_n$ 为简单样本（总体分布不必连续）时，有

$$(\widetilde{L}_n-n\bar{c}_n\bar{a}_n)/\sigma_n(\tau)\xrightarrow{\mathscr{L}}N(0,1)。\qquad (4\cdot30)$$

类似地，定理 4·5 和 4·6 也可以完全平行地推广到现在的情况。一言以蔽之，只要定理 4·4—4·6 中某一个的条件成立，而线性秩统计量 L_n 按平均法修改为 $\widetilde{L}_n$，则相应于该定理的渐近正态结果对 $\widetilde{L}_n$ 有效。

根据这一结果，如使用平均法，则在样本大小较大的情况下，使用极限分布作检验并无困难。

四、样本独立但不同分布时

前面几段所研究的都是样本为独立同分布的情况。在秩方法的研究中，也需要考虑样本不同分布的情形。例如，在两样本问题中，在原假设下两总体分布 F、G 相同，样本固然为独立同分布。但如要研究检验的功效，则涉及对立假设，因而要考虑 F 与 G 不同的情形。这时样本只独立但不同分布。

当样本不同分布时，无论是线性秩统计量的精确分布或其极限分布问题，都比同分布时复杂得多。就两样本问题（这时只涉及两个不同的分布，情况相对而言简单些）的线性秩统计量来说，第一个有普遍意义的重要结果是 Chernoff 和 Savage 在 1958 年作出的，后来到 1965 年经过 Govindarajulu 作了改进，以下我们将不加证明地引述这结果的一个特殊情况。到 1968 年，Hajek 在一项重要工作中，讨论了各样本的分布都可以不同的情况。

现设 $X_1,\cdots,X_{n_1}$ 和 $Y_1,\cdots,Y_{n_2}$ 分别为抽自具分布 F 和 G 的总体的简单样本，且假定合样本全体独立，又 F 和 G 都处处连续。记 $n=n_1+n_2$。以 R_i 记 Y_i 在合样本中的秩，$i=1,\cdots,n_2$。设函数 $a(u)$ 定义于 $0<u<1$。令

$$S_n=\frac{1}{n_2}\sum_{i=1}^{n_2}a\left(\frac{R_i}{n+1}\right). \tag{4·31}$$

事实上，此统计量与 n_1 和 n_2 都有关，理应记为 $S_{n_1n_2}$。为简便计就记为 S_n。以下几个量也属于这种情况：以 $H_n(x)$ 记合样本 $X_1,\cdots,X_{n_1}$，$Y_1,\cdots,Y_{n_2}$ 的经验分布函数，$\lambda_n=n_2/n$，而

$$\mu_n=\int_{-\infty}^{\infty}a\left(H_n(x)\right)dG(x), \tag{4·32}$$

$$\sigma_n^2=2(1-\lambda_n)\left(\sigma_{n2}^2+\frac{1-\lambda}{\lambda}\sigma_{n_1}^2\right), \tag{4·33}$$

其中

$$\sigma_{n1}^2 = \iint_{-\infty<x<y<\infty} F(x)(1-F(y))a'(H_n(x))a'(H_n(y)),$$
$$\cdot dG(x)dG(y), \tag{4·34}$$

$$\sigma_{n2}^2 = \iint_{-\infty<x<y<\infty} G(x)(1-G(y))a'(H_n(x))a'(H_n(y))$$
$$dF(x)dF(y), \tag{4·35}$$

有如下的定理：

定理 4·7 在上述诸假定和记号下，再假定：

(1) 存在$\varepsilon_1>0$，使 $\varepsilon_1<\lambda_n<1-\varepsilon_1$，对一切$n$。

(2) 存在常数 $\delta>0$ 及K，使

$|a^{(i)}(u)|\leqslant K(u(1-u))^{-i-1/2+\delta}$，$0<u<1$，$i=0,1$

此处 $a^{(0)}(u)=a(u)$，$a^{(1)}(u)=a'(u)$。

(3) 存在 $\varepsilon_2>0$，使

$$\max(\sigma_{n_1}^2,\sigma_{n_2}^2)\geqslant\varepsilon_2,\ 对一切\ n, \tag{4·36}$$

则当 $n\to\infty$时有

$$\sqrt{n}\ (S_n-\mu_n)\ /\sigma_n\xrightarrow{\mathscr{L}}N(0,1). \tag{4·37}$$

类似的结论对定理 4·5 中考虑的那种计分也成立：设 $\xi_{(1)}\leqslant\cdots\leqslant\xi_{(n)}$ 是从分布函数 $A(x)$ 中抽出的简单样本，而

$$S_n=\frac{1}{n_2}\sum_{i=1}^{n_2}E(\xi_{(R_i)}), \tag{4·38}$$

又以 $a(\cdot)$ 记 $A(\cdot)$ 的反函数。

定理4·8 若对这样定义的函数 $a(\cdot)$，定理 4·7 的条件(1)～(3)都满足，而 S_n 由(4·38)定义，则当 $n\to\infty$时仍成立(4·37)。

五、符号秩统计量

符号秩统计量来源于对称中心的检验问题．设 $X_1,\cdots,X_n$ 是从总体分布 $F(x-\theta)$中抽得的简单样本，其中 $F(x)$为关于原点对称的分布，θ 为实参数（即总体分布关于θ 对称），F，θ 都未知，要检验原假设 $\theta=\theta_0$，或 $\theta\leqslant\theta_0$（或$\theta\geqslant\theta_0$）．以下不失普遍

性，总设 $\theta_0=0$。这只要用 $X_i-\theta_0$ 代 X_i 即可。如要用秩方法来检验这个问题，则我们可这样想：设若 $\theta\neq0$，比如说 $\theta>0$。则样本 $X_1,\cdots,X_n$ 中，取正值者倾向于多，而那些取正值的样本，其在 $\{|X_1|,|X_2|,\cdots,|X_n|\}$ 中的秩也倾向于大。基于以上的考虑，可建立原假设的种种秩检验法，细节待以后再讲，此刻我们只注意：在上述考虑中，既涉及 $|X_i|$ 在 $\{|X_1|,\cdots,|X_n|\}$ 中的秩，也涉及 X_i 的符号。这导致以下关于符号秩的概念。

定义 4·2 暂设 $|X_1|,\cdots,|X_n|$ 互不相同。记

$\psi_i=I_{(X_i>0)}$，$R_i^+=|X_i|$ 在 $\{|X_1|,\cdots,|X_n|\}$ 中的秩，$i=1,\cdots,n$。则

$$R^+=(\psi_1R_1^+,\cdots,\psi_nR_n^+), \tag{4·39}$$

称为样本 $X_1,\cdots,X_n$ 的符号秩统计量。

简言之，符号秩统计量的意思是：若 $X_i\leqslant0$，则其符号秩定为 0。若 $X_i>0$，则 X_i 之符号秩定义为 X_i 在 $\{|X_1|,\cdots,|X_n|\}$ 中之秩。任何由 R^+ 派生出的统计量也称为符号秩统计量，例如

$$W^+=\sum_{i=1}^n\psi_iR_i^+ \quad \text{（Wilcoxon一样本符号秩和）}。 \tag{4·40}$$

这是一线性符号秩统计量。更一般的形式为

$$L_n^+=\sum_{i=1}^na(R_i^+)\psi_i, \tag{4·41}$$

此处 $a(j)$ 定义在 $j=1,2,\cdots,n$ 上。

关于符号秩统计量的分布有如下的结果：

定理 4·9 设 $X_1,\cdots,X_n$ 为抽自总体分布 F 的简单样本，F 处处连续且关于 0 对称。定义 ψ_i,R_i^+ 如前，则 $(\psi_1,\cdots,\psi_n,R_1^+,\cdots,R_n^+)$ 的联合分布由以下几条所决定：

(1) $(\psi_1,\cdots,\psi_n)$ 及 $(R_1^+,\cdots,R_n^+)$ 独立，且 $\psi_1,\cdots,\psi_n$ 为 iid.；

(2) $P(\psi_i=1)=P(\psi_i=0)=1/2$，$i=1,\cdots,n$；

(3) $(R_1^+,\cdots,R_n^+)$ 取 $(1,\cdots,n)$ 的任一置换的概率都是 $1/n!$。

证明 每个 ψ_i 只能取 0,1 两值。故 $(\psi_1,\cdots,\psi_n)$ 只能取形如

$(\eta_1,\cdots,\eta_n)$的 2^n 个值，其中 η_i 为 0 或 1.任意固定这样一个值，例如 $(0,\cdots 0,1,\cdots,1)$（前面m个为0，后 $n-m$个为 1）。在条件 $(\psi_1,\cdots,\psi_n)=(0,\cdots 0,1,\cdots,1)$ 之下考虑 $(R_1^+,\cdots,R_n^+)$ 的条件分布。由所设条件知，$X_1,\cdots,X_m$ 取非正值，而 $X_{m+1},\cdots,X_n$ 取正值.由于X关于 0 对称，知对每个 X_i,在 $X_i>0$ 的条件下$|X_i|$的条件分布，与在条件 $X_i<0$ 之下 $|X_i|$ 的条件分布一样（由于F连续，一点 0 处的概率为 0，可以不计），事实上,易知这两个情况下条件分布函数都是$\tilde{F}(x)=(2F(x)-1)I_{(x>0)}$.由于$\tilde{F}$ 处处连续（此由F处处连续，及F关于 0 对称因而 $F(0)=1/2$ 可知），由定理 4·1，知$(R_1^+,\cdots,R_n^+)$之分布如本定理的（3）所示.以上的推理在$(\psi_1,\cdots,\psi_n)$固定为上述个 2^n 个值中任何一个都对.这样一来，$(R_1^+,\cdots,R_n^+)$的条件分布与$(\psi_1,\cdots,\psi_n)$取的值无关，因而证明了二者独立，且（$R_1^+,\cdots,R_n^+$）的无条件分布，就与上述条件分布相同。这证明了本定理的(1)和(3)至于(2)，它是 X_1，…，X_n独立且 X_i 的分布关于 0 对称的简单推论.定理证毕.

利用这个定理，原则上可以（在定理条件下）定出符号秩统计量 R^+ 及任何线性符号秩统计量的分布.这一分布可供在 n 不大时，检验原假设 $\theta=0$ 或 $\theta\leqslant 0$ 之用.例如，设 $n=4$.若定理4·9的条件适合且 $\theta=0$，则易算出(具体计算留给读者)，由（4·40）定义的 Wilcoxon 一样本符号秩和统计量 W^+ 有分布当下：

$$P(W^+=i)=\frac{1}{16},\quad i=0,1,2,8,9,10;$$

$$=\frac{2}{16},\quad i=3,4,5,6,7.$$

设要检验的原假设为 $\theta=0$，对立假设为 $\theta\neq 0$.当$\theta\neq 0$ 时,W^+ 倾向于走向两个极端（取大值（当 $\theta>0$）或小值（当 $\theta<0$）），故应取边上之值置于否定域中，如取检验水平 $\alpha=1/8$，可取否定域为$\{0,10\}$.若 $\alpha=1/4$,可取否定域为$\{0,1,9,10\}$.如原假设为 $\theta\leqslant$（对立假设$\theta>0$），则应只取 W^+ 的大值于否定域。如 $\alpha=1/8$，

可取否定域为{9,10}。当 α 不是 1/16 的倍数时，如要严格达到预定的水平 α，则必须施行随机化。当 n 较大时，这往往没有必要，因可以通过稍微调整 α 之值，以避免这种随机化。

如果 n 相当大，则 W^+，或其他线性符号秩统计量，其分布过于复杂不便应用。这时可使用极限分布。

考虑一串线性符号秩统计量 $L_n^+=\sum_{i=1}^{n}a_n(R_i^+)\Psi_i$。记

$$\bar{a}_n=\frac{1}{n}\sum_{i=1}^{n}a_n(i),\qquad A_n^2=\frac{1}{n}\sum_{i=1}^{n}a_n^2(i).$$

定理 4·10 设样本 $X_1,\cdots,X_n$ 满足定理 4·9 的条件，且 $\{(a_n(1),\cdots,a_n(n)):n=1,2,\cdots\}$ 满足条件 N。则当 $n\to\infty$ 时，有

$$2\left(L_n^+-\frac{n}{2}\bar{a}_n\right)/\sqrt{nA_n^2}\xrightarrow{\mathscr{L}}N(0,1).\qquad(4\cdot42)$$

证明 根据定理 4·9 易知，L_n^+ 与 $\sum_{i=1}^{n}a_n(i)\Psi_i$ 同分布。故为证(4·42)，不妨设 L_n^+ 就是 $\sum_{i=1}^{n}a_n(i)\Psi_i$。由于 $\Psi_1,\cdots,\Psi_n$ 独立，这表达式是一个独立和，其渐近正态性可用中心极限定理去处理。此处我们使用 ЛЯПУНОВ 定理。由于 $E\Psi_i=1/2$，$\mathrm{Var}\Psi_i=1/4$，根据该定理，为证(4·42)，只须证明

$$\lim_{n\to\infty}\sum_{i=1}^{n}|a_n(i)|^3E\left|\Psi_i-\frac{1}{2}\right|^3\Big/\left(\frac{1}{4}nA_n^2\right)^{3/2}=0,\qquad(4\cdot43)$$

而此式显然可由下式推出：

$$\lim_{n\to\infty}\max_{1\le i\le n}|a_n(i)|\Big/\left(\sum_{i=1}^{n}a_n^2(i)\right)^{1/2}=0.\qquad(4\cdot44)$$

注意到

$$\max_{1\le i\le n}|a_n(i)|\le\max_{1\le i\le n}|a_n(i)-\bar{a}_n|+|\bar{a}_n|$$

及

$$\sum_{i=1}^{n} a_n^2(i) = \sum_{i=1}^{n}(a_n(i)-\bar{a}_n)^2 + n\bar{a}_n^2,$$

(4·44)是 $\{(a_n(1),\cdots,a_n(n)): n=1,2,\cdots\}$ 满足条件N的简单推论，于是证明了(4·43)，因而定理 4·10。

应注意到本定理与定理 4·4 的实质不同处。本定理处理的统计量貌似复杂，实际上即为通常的独立和，在这一点上说没有什么新东西。定理 4·4 中的线性秩统计量不是独立和，其处理用到特殊的技巧。

定理4·10有两个特例值得注意：

1. $a_n(i)=\varphi\left(\frac{i}{n+1}\right)$，而 $\varphi\in SS$（见定理 4·4 前面一段的说明）。

2. $a_n(i)=E\varphi(U_{ni})$，$\varphi\in SS$。这里$U_{n1}\leqslant\cdots\leqslant U_{nn}$是从(0,1)均匀分布中抽出的次序样本。

这只要证明，如此定义的$a_n(i)$使条件N满足。对前者这很容易，留作为习题。后一条的证明则比较难一些。

例 4·5 对 Wilcoxon 符号秩和检验 W^+，有 $a(i)=i$，算出

$$\bar{a}_n=(n+1)/2,$$

$$A_n^2=\frac{1}{6}n(n+1)(2n+1)/n=\frac{1}{6}(n+1)(2n+1),$$

据(4·42)，得到双侧假设 $\theta=0$ 的水平 α 大样本否定域为

$$\left\{\left|L_n^+-\frac{n(n+1)}{4}\right|>\left((n(n+1)(2n+1))^{1/2}/(2\sqrt{6})\right)u_{\alpha/2},\right.$$

单侧假设 $\theta\leqslant 0$ 的否定域则是

$$\left\{W^+>\frac{n(n+1)}{4}+\left(\left(n(n+1)(2n+1)\right)^{1/2}/(2\sqrt{6})\right)u_\alpha\right\}.$$

若取 $a(i)=1$，所得检验B称为符号检验。对此检验有

$$\bar{a}=1,\qquad A_n^2=1.$$

按(4·42)，得到双侧假设 $\theta=0$ 的水平 α 大样本否定域是 $\{|B-$

$\frac{n}{2}| > \frac{1}{2}\sqrt{n}u_{\alpha/2}\}$．单侧假设 $\theta\leqslant 0$ 的否定域则是

$$\{B>\frac{n}{2}+\frac{1}{2}\sqrt{n}u_{\alpha}\}.$$

§4·2 一、两样本检验及其优良性

如前所述，一样本问题是指：设从一对称分布中抽出了一些简单样本，要据以检验关于对称中心 θ 的假设，一般是 $\theta=\theta_0$、$\theta\leqslant\theta_0$ 及 $\theta\geqslant\theta_0$ 等，θ_0 为给定的数。两样本问题则指有两组样本分别抽自分布 F 和 G，要检验假设 $F=G$，或其他单侧性的假设。例如已知 $G(x)=F(x-\theta)$，要检验 $\theta\leqslant 0$。这些问题在实用上有重大意义。在参数统计中，往往假定总体分布为正态型，这时常用的检验法就是熟知的一、两样本 t 检验。在对总体分布并无特定的假定时，问题为非参数性的。统计学者提出了许多检验法，使用线性秩统计量及线性符号秩统计量的方法，是其中重要的一类。

本节将先提出这类检验法中一些著名的例子。通过这些例子看到，同一问题可供选择的秩检验很多。那么选那一个好呢？要回答这个问题，就需要考察秩检验的优良性。本节将提出两种优良性准则，作为比较的标准。

一、重要的一、两样本秩检验

1．两样本位置参数秩检验

问题的提法，我们在本书的开篇处的例 1·1 中就表述过了：简单样本 $X_1,\cdots,X_{n_1}$ 来自分布 $F(x)$，而 $Y_1,\cdots,Y_{n_2}$ 来自 $F(x-\theta)$。分布 F 及参数 θ 都未知，要检验关于 θ 的假设，通常有

$$H_1:\theta=\theta_0;\qquad H_2:\theta\leqslant\theta_0;\qquad H_3:\theta\geqslant\theta_0,$$

各有相应的对立假设。不失普遍性以下总假定 $\theta_0=0$。又假定 F 处处连续，如这个不成立，则用处理结的方法去对付。

用秩方法来检验这些假设，思想很简单：以 $R_1,\cdots,R_{n_2}$ 分别

记$Y_1,\cdots,Y_{n_2}$在合样本中的秩。若$\theta>0$，则因每个Y_i的分布与每个$X_i+\theta$的分布相同，Y样本倾向于取比X样本更大的值。注意这里"倾向于"是一种统观而含糊的说法，它并不意味着Y样本一定比X样本大。可以理解为：Y样本大于X的样本的"机会"更多，而小于它的机会则少。这样一来，$R_1,\cdots,R_{n_2}$当$\theta>0$时倾向于取集合$\{1,2,\cdots,n\}$中较大的值$(n=n_1+n_2)$。同样，若$\theta<0$，则$R_1,\cdots,R_{n_2}$倾向于取集合$\{1,2,\cdots,n\}$中较小的值。因此，若取一个定义在$\{1,2,\cdots,n\}$上非降的计分函数$a(\cdot)$，则统计量

$$L=\sum_{i=1}^{n_2}a(R_i), \tag{4·45}$$

当$\theta>0(\theta<0)$时倾向于取大（小）值。因此，在检验原假设H_1时，可以把L的两端的极端值放入否定域。如果$n_1=n_2$，或者函数a满足条件（4·7），则根据定理4·2，当原假设H_1成立时L的分布关于点$n_2\bar{a}$对称，因而否定域可取为

$$\{|L-n_2\bar{a}|>C\}\ \left(\bar{a}=\frac{1}{n}\sum_{i=1}^{n}a(i)\right). \tag{4·46}$$

常数C根据水平α定。如果计分函数$a(\cdot)$满足定理4·4或4·5中的条件，且$\min(n_1,n_2)\to\infty$，则在H_1成立时，有

$$\sqrt{\frac{n(n-1)}{n_1n_2}}\,(L-n_2\bar{a})\Big/\sqrt{\sum_{i=1}^{n}(a(i)-\bar{a})^2}\xrightarrow{\mathscr{L}}N(0,1). \tag{4·47}$$

因此当n_1和n_2都较大时，对给定的水平α，（4·46）中的C近似地可取为

$$C=\left(\sum_{i=1}^{n}(a(i)-\bar{a})^2\frac{n_1n_2}{n(n-1)}\right)^{1/2}u_{\alpha/2}. \tag{4·48}$$

选择种种适合上述条件的计分函数$a(\cdot)$，就可以作出种种不同的秩检验，其中有几个著名的我们已在前面提到过：

Wilcoxon检验：取$a(i)=i$；

Fisher-Yates检验：取$a(i)=E\xi_{ni}$，其中$\xi_{n1}\leqslant\cdots\leqslant\xi_{nn}$是从标准正态分布$N(0,1)$中抽出的次序样本（前已指出，Fisher-

Yates 表中载有 $E\xi_{ni}$ 之值）。注意对这个 $a(\cdot)$ 有 $\bar{a}=0$。Terry 在 1952 年讨论过这个检验，故有时这检验也冠以 Fisher-Yates-Terry 之名称。

Van der Waerden 检验取 $a(i)=\Phi^{-1}\left(\frac{i}{n+1}\right)$，此处 Φ^{-1} 为标准正态分布函数 Φ 的反函数，这个检验是 Van der Waerden 在 1952年提出的。

之所以要考虑种种不同的检验，其理由正如治同一种病有若干种方法，其选用根据具体情况而定。在此，总体分布 F 如何，与检验的性能有很大关系。针对在应用中可能遇到的种种 F 设计不同的检验，在使用时根据所了解的情况从中适当挑选，就能达到更好的效果。例如，若 F 为正态，则以后将指明：在各种秩检验中以采用 Fisher-Yates 检验或 Van der Waerden 检验最好。读者可能会有这样的问题：如 F 为正态，则 t 检验是一久经考验的优良检验，何必还要用 Fisher-Yates 或其他秩检验？这问题问得好。问题在于：我们可能有相当的把握认为 F 是正态，而无确实的把握。若仅采用 t 检验，则万一 F 真不为正态，就可能产生严重后果。为防备这种可能，我们采用非参数提法，而又把效率最大的方向定在正态分布上，以使当分布 F 确为正态时效果很好，而即使 F 不为正态，检验仍维持一定的性能。这样就基本上兼顾了两方面的需要。

以上的分析也说明了这样一个重要的思想：虽则非参数统计方法是建立在模型很广的基础上，以对付由于对模型分布无确切了解的情况，但这决不等于说，当使用非参数方法时，我们可以不用费力去搜集关于总体分布的尽可能充分的知识。相反，这步工作做得愈好，我们对总体分布了解得愈多，就愈有可能选择能针对当前问题的方法，对总体分布的了解，除依据对问题的专业知识，有关的理论及以往的经验外，样本数据也常能提供一些有用的信息。

以上的讨论针对假设 $\theta=0$。对单侧假设 $\theta\leqslant0$ 或 $\theta\geqslant0$，情况完全相似。如对 $\theta\leqslant0$，否定域应取为 $L<C^*$。按渐近正态近似，C^* 可取为

$$C^*=n_2\bar{a}+\left(\sum_{i=1}^{n}(a(i)-\bar{a})^2\frac{n_1n_2}{n(n-1)}\right)^{1/2}u_\alpha \qquad (4\cdot49)$$

2．两样本刻度参数秩检验

设有抽自总体分布 $F(x)$ 的简单样本 $X_1,\cdots,X_{n_1}$ 和抽自总体分布 $F(x/\sigma)$ 的简单样本 $Y_1,\cdots,Y_{n_2}$。此处分布 F 及参数 $\sigma>0$ 都未知。要检验关于 σ 的假设，通常有 $H_1:\sigma=\sigma_0$，$H_2:\sigma\leqslant\sigma_0$ 和 $H_3:\sigma\geqslant\sigma_0$，各有相应的对立假设。不失普遍性以下总假定 $\sigma_0=1$。这可以通过用 σ_0X_i 代替 X_i 而达到。

σ 称为刻度参数，是因为 Y_i 的分布与 σX_i 相同，换句话说，从分布的角度看，X 样本与 Y 样本之差别，相当于同一个量在不同单位的坐标系之下所产生的差别。

刻度参数两样本问题比位置参数的情况要复杂些。问题在于，在位置参数的情况，Y 样本（就分布而言）相当于 X 样本加上 θ。因此 $\theta>0$ 时 Y 样本秩倾向大，$\theta<0$ 时倾向小，这个总的趋势与总体分布 F 无关（这一点很重要）。但在刻度参数情况则不然。例如，设 $\sigma>1$。这时，Y 样本（在分布上）相当于 X 样本乘以 σ。如果 X 的分布 F 全在正轴一边（即 $F(0)=0$），这时乘以 σ 的后果使 Y 样本倾向于增大。反之，若 F 全在负轴一边（$F(0)=1$），则情况正好相反。若 F 在原点两边都有分布，则乘以 $\sigma>1$ 的后果是使正者更大，负者更小，因而 Y 样本倾向于走极端。究竟是那种情况，需要有关 F 的知识。

解决这个困难的途径有二．一是在总体分布 F 上附加一定的条件。例如，F 关于原点对称，或至少其中位数为0。这时，落在0两边的样本个数大致相当。故乘以 $\sigma>1$ 后，走向两边极端的样本个数也大致相当。这个考虑引导到下述秩检验方法：选定一个"两头大中间小"的计分函数 $a(\cdot)$，即满足条件

$$a(1) \geqslant a(2) \geqslant \cdots \geqslant a\left(\left[\frac{n+1}{2}\right]\right) \leqslant a\left(\left[\frac{n+1}{2}\right]+1\right) \leqslant \cdots \leqslant a(n). \quad (4\cdot50)$$

此处 $n=n_1+n_2$，而 $\left[\frac{n+1}{2}\right]$为不超过$\frac{n+1}{2}$的最大整数。设原假设为 $\sigma\leqslant1$。当对立假设成立时 $\sigma>1$.这时，$Y_1,\cdots,Y_{n_2}$ 在合样本中的秩倾向于跑到两侧极端。按（4·50），$L=\sum_{i=1}^{n_2} a(R_i)$ 会倾向于大.由此得出，应取 $L>C$ 为否定域，C 可根据水平 α，通过小样本分布（当 n 小时）定出，或在 n 较大时，用正态逼近而得到的公式（4·49）定出.如果检验 $\sigma=1$，则否定域取为双侧的，在大样本情况由（4·46）和（4·48）定出。

其所以要假定中位数为 0，有重要的理由。设想一种情况：$F(0)=1/5$.这时，有 4/5 的概率使X样本取正值，1/5取负值，故大体上说，X样本和Y样本各有 4/5 左右在 0 点右边。这部分样本Y倾向于比X大。另 1/5 左右的样本则相反：Y 倾向于小于X，从整体而言，$\sigma>1$ 使Y样本之秩增大者多而减小者小，这时，一个单调上升的计分函数要比满足（4·50）的计分函数更切合于检验 $\sigma\leqslant1$，而（4·50）的分辨力很差.只有在中位数为0时，0两边样本个数相当，选择满足条件（4·50）的计分函数才最有利.几个有名的例子如下：

$$\text{Mood 检验：} a(i)=\left(i-\frac{n+1}{2}\right)^2$$

是Mood 在 1954 年提出的。

Ansary-Bradley 检验：$a(i)=\left|i-\frac{n+1}{2}\right|$，是这两位作者在 1960 年提出的。

Copan 检验：$a(i)=E\xi_{ni}^2$，此处 $\xi_{n1}\leqslant\cdots\leqslant\xi_{nn}$ 是从 $N(0,1)$ 中抽出的次序样本。注意，由 $N(0,1)$ 关于 0 对称易知（请读者写出证明）（4·50）满足，且 $a(i)=a(n+1-i)$.此检验是Copan

在 1961 年提出的。

Klotz 检验：$a(i)=\left(\Phi^{-1}(\frac{i}{n+1})\right)^2$，1962年提出。此处 Φ 为 $N(0,1)$ 之分布函数，Φ^{-1} 为其反函数。我们留给读者去验证：（4·50）满足，且 $a(i)=a(n+1-i)$。

Siegel-Tukey 检验：是 1961 年提出的。$a(i)$ 的取法是：$a(1)=n$，$a(n)=n-1$，$a(n-1)=n-2$，$a(2)=n-3$，$a(3)=n-4$，$a(n-2)=n-5$，$a(n-3)=n-6$，$a(4)=n-7$，…，读者不难看出其一般规律何在。

根据定理 4·4 和 4·5，不难验证，在以上所有的检验中，$a(\cdot)$ 的取法都使在 $\sigma=1$（即两总体同分布）之下，检验统计量 $L=\sum_{i=1}^{n}a(R_i)$ 适合渐近正态定理（4·47）。其中，Siegel-Tukey 检验中 $a(\cdot)$ 的取法有一个特点，即 $a(i)$，$i=1,\cdots,n$，取遍 $1,2,\cdots,n$ 中各值且仅取一次。因此，在 $\sigma=1$ 之下，该检验的检验统计量与熟知的 Wilcoxon 统计量有同一分布。

另一种作法是把样本作一个平移，以达到中位数为0的要求。具体作法是：用 X 样本对 $F(x)$ 的中位数作一估计。设为 $\hat{m}_1$（例如，取 $X_1,\cdots,X_{n_1}$ 的样本中位数）。类似地，用 Y 样本对 $F(x/\sigma)$ 的中位数作一估计，设为 $\hat{m}_2$。然后，令

$$X_i'=X_i-\hat{m}_1,\ i=1,\cdots,n_1;\quad Y_j'=Y_j-\hat{m}_2,\ j=1\cdots,m_2,\tag{4·51}$$

再从 $(X_1',\cdots,X_{n_1}')$ 和 $(Y',\cdots,Y_{n_2}')$ 出发，按前面已知中位数为 0 的情况去处理即可。应当注意的是：由于 $\hat{m}_1$ 依赖于所有的 $X_1,\cdots X_{n_1}$，故 $X_1',\cdots,X_{n_1}'$ 将不再是独立的。对 $Y_1',\cdots,Y_{n_2}'$ 也有同样的问题，因此，形式上再引用定理 4·4 和 4·5 已不行，必须再加上补充的论证。当 n_1 和 n_2 都相当大时，估计量 $\hat{m}_1$ 和 $\hat{m}_2$ 与中位数真确值 m_1 及 m_2 很接近。因此这种做法，与中位数严格为 0 的情况相比，不会有多大的偏差。n_1 和 n_2 较小时则不然，且作过变换（4·51）后，L 在 $\sigma=1$ 之下的确切分布也很不易求，

其所以要提出一些检验法供选择，其理由当然与位置参数的情况相同。

3．对称中心的检验

问题提法已在§4·1的五段中说明过了，样本$X_1,\cdots,X_n$抽自总体分布$F(x-\theta)$，已知分布$F(x)$关于0对称，因而θ为总体分布的对称中心，要检验假设$\theta=\theta_0$，$\theta\leqslant\theta_0$或$\theta\geqslant\theta_0$，各有相应的对立假设；不失普遍性可设$\theta_0=0$。

检验方法在§4·1的五段中已说明过：先由样本定出符号秩，再选定计分函数$a(\cdot)$而作出形如(4·41)的检验统计量L_n^+。$a(\cdot)$必须在集$\{1,2,\cdots,n\}$上非降，且使条件N满足。这时当n较大时，可以定理4·10去决定检验统计量的临界值。

$a(\cdot)$的不同取法导致种种的检验法，以适应各种不同的情况。最重要的除前述的Wilcoxon一样本符号秩和检验（见(4·40)）外，还有符号检验(Signtest)，相当于取$a(i)=1$，$i=1,\cdots n$。由它产生的检验统计量，就是样本$X_1,\cdots,X_n$中符号为正的个数。符号检验以此得名。Fisher-Yates和Van der Waerden检验，也可移至一样本情况，此处不细述了。

二、检验的渐近相对效率

在一段中，就几个重要的检验问题，提出了一些秩检验，以备在种种可能的模型下去选择使用。为具体进行挑选，就需要确定一种准则。不同的检验在该准则下比较其优劣，而决定去取。

这里从两个检验的相对效率这个角度，来引进一种比较的准则。简言之，设为检验同一假设，有A、B两个检验可用。定义A对B的相对效率，暂记为$e_{A,B}(F)$。这里的F表示模型中涉及的分布。设想我们对所有可能的F都能计算$e_{A,B}(F)$之值。对当前的检验问题，我们先根据所了解的情况，确定一个认为最可能的F，比方说，确定为正态分布。对F为正态去计算$e_{A,B}(F)$之值。若它大于1，则A优于B而我们选择检验A。若$e_{A,B}(F)<1$则选择B。有时我们对F的了解不多，不足以有把握地确定一个可能的

F. 即使在这种情况，这个方法仍有参考意义. 比方说，我们对 $e_{A,B}(F)$ 当 F 取种种典型的分布时其取值情况有了解，则会发现，对多数在实用上有意义的分布而言，$e_{A,B}(F)$ 大于 1 者居多. 这可以解释为 综合地看 A 优于 B. 在对 F 知之不多的情况下，选 A 的理由就显得更充足.

故使用这个方法有两个步骤. 首先是 $e_{A,B}(F)$ 如何定义，其次是如何对具体的 F 算出其数值. 在相当大程度上可说，后一步包含在前一步之内，因为有了定义，就能导出其表达式. 实际计算时当然有可能要用到数值方法.

根据假设检验中流行的 Neyman-Pearson 理论，同一水平下的两个检验，功效大者为优. 也可以换一个说法：在同一水平之下，为在同一的对立假设下达到同一功效，需要样本少者为优. 且样本大小之反比可定为相对效率.

例如，用 Wilcoxon 检验（以下简称 W 检验）去检验两样本位置参数 θ 为0. 取 $n_1=n_2=5$，水平 $\alpha=\frac{2}{63}$. 实际计算表明：若总体分布为方差 1 的正态分布，而位置参数 θ 之真值为 0.5、1 和 1.5（原假设不成立）时，W 检验的功效分别为 0.072，0.210和 0.431. 而为要 t 检验在同一检验水平(2/63)之下，在上述 θ 值处达到同一功效，分别需取 $n_1=n_2$ 之值为：4·840，4·890，4·805（t 检验的功效是由非中心 t 分布计算的，其表达式在样本大小非整数时也有意义）. 由此算出，在所述情况下，W 检验对 t 检验的相对效率，分别为

$$\frac{4.840}{5}=0.9680,\quad \frac{4.890}{5}=0.9780,\quad \frac{4.805}{5}=0.9610.$$

这说明 t 检验（至少在所述情况下）优于 W 检验. 但也许更有兴趣的是：即使在 t 检验最能发挥优势的场合（t 检验本来就是针对正态分布设计的），且在样本大小如此小的场合，t 检验对 W 检验的优势其实也很小.

从这个例子也看出，相对效率是一个很复杂的东西。除了模型中的分布外，它还依赖于检验的水平，样本大小，以及对立假设的位置。这样一个复杂的量难以掌握，我们希望能加以简化，使它只依赖于模型的分布 F，而与后面三个因素无关。Pitman 在1948年通过取极限的步骤，定义了一个具有这种性质的量。这就是所谓"渐近相对效率"（Asymptotic Relative Efficency，简记为ARE）。不难理解，这种简化要付出一定的代价，即它只有在相当严格的条件下，并在相当局限的范围内，才可实现。下面就来着手定义 Pitman 的 ARE。

设有样本 $Z_1,\cdots,Z_n$（如在两样本问题，$Z_1,\cdots,Z_n$ 可以是 $X_1,\cdots,X_{n_1}$, $Y_1,\cdots,Y_{n_2}$），其分布依赖于某分布 F 及一个实参数 θ。在前面考察的一、两样本问题都是这个情况。考虑假设检验问题

$$\theta=\theta_0\longleftrightarrow\theta>\theta_0, \tag{4·52}$$

此处为方便计，将原假设取为 $\theta=\theta_0$。原假设为 $\theta\leqslant\theta_0$ 的情况只须作少许修改。设 S 和 T 是这问题的两个检验。当样本大小为 n 时，S 和 T 可更明确地记为 S_n 和 T_n（故在此 S 和 T 不过是检验的一个名称而已，如 t 检验，秩和检验，符号检验之类），S_n 和 T_n 也拿来记检验统计量，而检验的否定域分别为 $\{S_n>c_n\}$ 以及 $\{T_n>d_n\}$。分别以 $\beta_S(n,F,\theta)$ 和 $\beta_T(n,F,\theta)$ 记检验 S_n 和 T_n 的功效函数。

定义4·3 （Pitman 的 ARE） 设对任何指定的 $\alpha,\beta,0<\alpha<\beta<1$，及一串下降趋于 θ_0 的对立假设值 $\theta_k\downarrow\theta_0$，可找到自然数的两个子列 $\{n_k\}$ 及 $\{n_k'\}$，以及检验统计量的临界值 c_n 和 d_n，满足以下的条件：

(1) $\lim\limits_{k\to\infty}\beta_S(n_k,F,\theta_0)=\alpha$, $\lim\limits_{k\to\infty}\beta_T(n_k',F,\theta_0)=\alpha$;

(2) $\lim\limits_{k\to\infty}\beta_S(n_k,F,\theta_k)=\beta$, $\lim\limits_{k\to\infty}\beta_T(n_k',F,\theta_k)=\beta$;

(3) 对任何满足(1)和(2)的序列$\{n_k'\}$及$\{n_k'\}$，极限$\lim_{k\to\infty} n_k'/n_k$必存在，其值与$\{n_k\}$及$\{n_k'\}$的取法无关，而且也与$\alpha,\beta$，$\theta_0$及收敛于$\theta_0$的序列$\{\theta_k\}$无关，则这极限定义为检验$S$对$T$的渐近相对效率，记为ARE$(S,T;\ F)$，我们来对定义中的几个条件作一些解释。

第(1)条是要求两检验S、T（或更确切地说，两序列检验$\{S_{n_k}\}$和$\{T_{n'_k}\}$，下同）都有渐近水平α。第(2)条是指在对立假设序列$\{\theta_k\}$上，二者的渐近功效相同。因为此处设了$\beta<1$，一般会有θ_k趋于θ_0。因为，若θ_k始终保持与θ_0有一定距离，而当样本大小很大时，通常在θ_k处的功效将趋于1（这一性质称为检验的相合性）。最后一条是为了使渐近效率与水平α、功效β及$\{\theta_k\}$都无关，而只成为S与T之间的对比（当然，它还依赖F，这正是所需要的），这样达到我们上面所说的简化。

至于对立假设选为由一实参数标定，是为了使在$k\to\infty$的过程中，对立假设的变化情况能更有规律些，以便使极限$\lim_{k\to\infty} n'_k/n_k$的存在成为可能。形象地可以这样设想：把原假设看成平面上的一个点，周围全是对立假设。用一实参数标定的对立假设类，相似于平面上由此点出发的一条半射线。在多元函数中，让变元随意地趋向一点，函数变化情况可以很复杂。但如沿一条半射线趋向该点，则转化为简单的一元情形。

看到这里，读者也许会感到仍有所不足：且不提定义中的那些限制性颇强的要求，这样定义的ARE(S,T)只在样本大小很大（理论上是无穷）时，才能成为比较二者效率的合理指标，而通常在实用中样本大小不一定非常大，因而在每一具体问题中，我们都难以仅凭借ARE去判定两检验何者为优。这问题提得好。可是，这不过只是统计学(及其他数学部门)中常用的一种做法，即在无法可施时转向取极限。这在相当程度上反映了本学科面貌的一个特点、水平及局限性，但也还有其它说法。使用电子计算

机等快速计算工具，在有限样本之下计算两检验的相对效率，并无原则困难。如果你在某一具体场合感到必须这样做，不会有人反对。但由于有限样本带来的复杂性，确实也可能掩盖某些本质的东西。通过取极限把它提取出来，无疑是很有认识意义的。即从实用的角度说，在许多重要情况下，由很小的样本算出的相对效率，已与 ARE 很接近。例如在前面讨论过的W检验和 t 检验的对比中，以下将证明对正态分布而言有 ARE$(W, t)=3/\pi=0.955$，这与那里算出的几个值极接近，而样本大小$n_1=n_2=5$

作了这些一般性的解释后，我们来在检验统计量有渐近正态性，及在另一些附加假定之下，推导出 ARE 的公式。施加的假定列举如下：

(1) 存在函数$\mu(S,n,\theta,F)$，$\mu(T,n,\theta,F)$，$\sigma(S,n,\theta,F)>0$及$\sigma(T,n,\sigma,F)>0$，使当$\{n_k\}$及$\{n'_k\}$满足定义 4·3 的(1)和(2)时，有

$$(S_{n_k}-\mu(S,n,\theta_0,F))/\sigma(S,n,\theta_0,F)\xrightarrow{\mathscr{L}}N(0,1),\quad(4\cdot53)$$

$$(T_{n'_k}-\mu(T,n'_k,\theta_0,F))/\sigma(T,n'_k,\theta_0,F)\xrightarrow{\mathscr{L}}N(0,1),\quad(4\cdot54)$$

$$(S_{n_k}-\mu(S,n_k,\theta_k,F))/\sigma(S,n_k,\theta_k,F)\xrightarrow{\mathscr{L}}N(0,1),\quad(4\cdot55)$$

$$(T_{n'_k}-\mu(T,n'_k,\theta_k,F))/\sigma(T,n'_k,\theta_k,F)\xrightarrow{\mathscr{L}}N(0,1).\quad(4\cdot56)$$

(2) μ 作为 θ 的函数，在 θ_0 附近可导，且当 $k\to\infty$时，有

$$\mu'(S,n_k,\theta_k,F)/\mu'(S,n_k,\theta_0,F)\to 1,\quad(4\cdot57)$$

$$\mu'(T,n'_k,\theta_k,F)/\mu'(T,n'_k,\theta_0,F)\to 1,\quad(4\cdot58)$$

(3) 当 $k\to\infty$时，有

$$\sigma(S,n_k,\theta_k,F)/\sigma(S,n_k,\theta_0,F)\to 1\quad(4\cdot59)$$

$$\sigma(T,n_k',\theta_k,F)/\sigma(T,n_k',\theta_0,F)\to 1. \tag{4·60}$$

(4) 当 $n\to\infty$ 时，有

$$\mu'(S,n,\theta_0,F)/(\sqrt{n}\,\sigma(S,n,\theta_0,F))\to K_S(F)\text{存在,} \tag{4·61}$$

$$\mu'(T,n,\theta_0,F)/(\sqrt{n}\,\sigma(T,n,\theta_0,F))\to K_T(F)\text{存在,} \tag{4·62}$$

且 $K_S(F)$ 和 $K_T(F)$ 不同为 0 和不同为 ∞。

在这些条件之下，注意到检验 S_{n_k} 和 $T_{n'_k}$ 的否定域分别为 $\{S_{n_k}>c_{n_k}\}$ 及 $\{T_{n'_k}>d_{n'_k}\}$，由定义 4·3 的(1)得

$$\lim_{k\to\infty}\frac{c_{n_k}-\mu(S,n_k,\theta_0,F)}{\sigma(S,n_k,\theta_0,F)}=\lim_{k\to\infty}\frac{d_{n'_k}-\mu(T,n_k',\theta_0,F)}{\sigma(T,n_k',\theta_0,F)}$$

$$=\Phi^{-1}(1-\alpha), \tag{4·63}$$

而由定义 4·3 的(2)得

$$\lim_{k\to\infty}\frac{c_{n_k}-\mu(S,n_k,\theta_k,F)}{\sigma(S,n_k,\theta_k,F)}=\lim_{k\to\infty}\frac{d_{n'_k}-\mu(T,n_k',\theta_k,F)}{\sigma(T,n_k',\theta_k,F)}$$

$$=\Phi^{-1}(1-\beta). \tag{4·64}$$

由(4·59)、(4·60)知，(4·64)中两个极限号下的表达式中的分母，分别可以用 $\sigma(S,n_k,\theta_0,F)$ 和 $\sigma(T,n'_k,\theta_0,F)$ 代替。代替后的式子与(4·63)相减，得

$$\lim_{k\to\infty}\frac{\mu(S,n_k,\theta_k,F)-\mu(S,n_k,\theta_0,F)}{\sigma(S,n_k,\theta_0,F)}$$

$$=\lim_{k\to\infty}\frac{\mu(T,n_k',\theta_k,F)-\mu(T,n_k',\theta_0,F)}{\sigma(T,n_k',\theta_0,F)}\neq 0.$$

将极限号下表达式的分子用中值定理，两边相除，利用(4·57)、(4·58)、(4·61)和(4·62)，即得

$$\lim_{k\to\infty}(\sqrt{n_k}K_S(F)/\sqrt{n'_k}K_T(F))=1,$$

因此

$$\lim_{k\to\infty}n_k'/n_k=K_S^2(F)/K_T^2(F)(=\text{ARE}(S,T;\ F)). \tag{4·65}$$

（4·65）右边不依赖于α、β、θ_0及$\{\theta_k\}$。因而适合定义4·3的条件，这样便证明了下面的定理。

定理4·11 在前述诸条件下，两检验S，T的渐近相对效率存在，且由（4·65）、（4·61）及（4·62）决定。

以上讲到的ARE的定义及计算公式，不止适用于秩检验。在每一具体实例中，都需要验证上面一大堆条件成立。就我们接触过的几种统计量（次序统计量，U统计量，秩统计量）而言，我们都曾证明或提到过其渐近正态定理。故从原则上说，我们已掌握足够的前提去验证这些条件，但这往往牵涉到无甚统计意义的繁琐细节。因此在下面讨论例子时，我们将不去严格地逐一地验证所有的条件。

从公式（4·65）看出，S对T的ARE为两个因子之比，其一只与S有关，另一只与T有关。因此，可以把$K_S^2(F)$称为检验S的效率因子。

上面只讨论了单侧假设的情况、双侧假设的情况，连同其导出的公式，都与此类似。

例4·6 $X_1,\cdots,X_n$为抽自$F(x-\theta)$的简单样本，$F(x)$关于0对称，且有密度函数$f(x)$。要检验假设：$\theta\leqslant 0$。考虑三个检验。其一是通常的t检验。其统计量为$T_n=\sqrt{n}\,\overline{X}_n/s_n$。$\overline{X}_n$为样本均值而$s_n^2$为样本方差。另两个是符号检验$B$，其统计量为$B_n=\sum\limits_{i=1}^{n}\psi_i$，$\psi_i=I(X_i>0)$，以及Wilcoxon符号秩和检验$W^+$其统计量为$W_n^+=\sum\limits_{i=1}^{n}\psi_i R_i^+$，后二者在§4·1五段中已讨论过。

为计算这些检验之间的ARE，只须算出各自的效率因子。按公式（4·61），这需要决定每个统计量的渐近正态形式中的函数μ和σ。

先考虑t检验。假定分布F的方差δ^2存在有限，则易见可取

$$\mu(t,n,\theta,F)=\sqrt{n}\,\theta/\delta,\quad \sigma(t,n,\theta,F)=1. \tag{4·66}$$

事实上，有

$$\sqrt{n}\,\bar{X}_n/s_n-\sqrt{n}\,\theta/\delta=\sqrt{n}(\bar{X}_n-\theta)/\delta-\sqrt{n}\,\bar{X}_n(s_n-\delta)/(s_n\delta). \qquad (4\cdot67)$$

由于 θ 和 δ^2 分别是总体分布 $F(x-\theta)$ 的数学期望和方差，由中心极限定理知，上式右边第一项依分布收敛于 $N(0,1)$。至于第二项，注意到 s_n 依概率收敛于 δ，即知该项依概率收敛于 0。这说明当 $n\to\infty$ 时，上式左边依分布收敛于 $N(0,1)$，从而证实了（4·66）中取法的可行性。

其次考虑 B_n。注意到 $\psi_1,\cdots,\psi_n$ 为 iid.，其数学期望为 $E\psi_1=P(X_1>0)=1-P(X_1\leqslant0)=1-F(x-\theta)|_{x=0}=1-F(-\theta)=F(\theta)$。最后一步用到 $F(x)$ 关于 0 对称。又 ψ_1 的方差为 $F(\theta)(1-F(\theta))$，当 $\theta\to0$ 时有极限 1/4。因此由中心极限定理知，可取

$$\mu(B,n,\theta,F)=nF(\theta),\quad \sigma(B,n,\theta,F)=\frac{1}{2}\sqrt{n}$$

（注意：条件（4·53）—（4·56）只要求（就本例而言），渐近正态性在 $\theta=0$ 及 $\theta\to0$ 时成立，并非要求固定的 θ 时成立。因此，本来按中心极限定理，应取 $\sigma(B,n,\theta,F)=(nF(\theta)(1-F(\theta)))^{1/2}$，但由于此式在 $\theta=0$ 或 $\theta\to0$ 时有极限

$$(n/4)^{1/2}=\frac{1}{2}\sqrt{n},$$

故直接取 $\sigma(B,n,\theta,F)=\frac{1}{2}\sqrt{n}$，这种取法使推导有所简化）。

最后考虑 W^+。这个情况比较费周折。因为，定理4·10只处理了在原假设下线性符号秩统计量的极限问题，而此处要求的比这多。幸好，统计量 W_n^+ 与 U 统计量有一个简单联系：引进核函数

$$h(x_1,x_2)=I(x_1+x_2>0), \qquad (4\cdot68)$$

并以 U_n 记以此为核的基于样本 $X_1,\cdots,X_n$ 的 U 统计量，则有

$$W_n^+=B_n+\binom{n}{2}U_n. \qquad (4\cdot69)$$

事实上，按 U_n 的定义，知

$$\binom{n}{2}U_n=\sum_{1\le i<j\le n}I(X_i+X_j>0). \tag{4·70}$$

我们把右边和号下所有为 1 的项分成 n 类 $C_1,\cdots,C_n$.其中类C_i包含一切这样的项：$X_i>0$，且 $|X_j|<X_i$（由于分布连续，可设$|X_1|,\cdots,|X_n|$互不相同）.不难见到：（4,70）和号下每个为1的项，必归入$C_1,\cdots,C_n$中之一类且仅一类.于是$\binom{n}{2}U_n$等于各类中包含之项之和.按符号秩的定义，C_i 类中之项恰为 $\psi_i R_i^+-\psi_i$.于是

$$\binom{n}{2}U_n=\sum_{i=1}^n(\psi_i R_i^+-\psi_i)=\sum_{i=1}^n\psi_i R_i^+-\sum_{i=1}^n\psi_i=W_n^+-B_n,$$

这证明了（4·69）.

对 U_n 可使用极限定理 3·1.为此要先算出 $\sigma_1^2(\theta)$，它是 $h_1(X_1)$的方差，其中

$$h_1(x)=Eh(x,X_2)=P(x+X_2>0)=P(X_2>-x)$$
$$=1-P(X_2\leqslant -x)=1-F(-x-\theta)=F(x+\theta).$$

如前，最后一步用了$F(x)$关于 0 对称的性质.有

$$\sigma_1^2(\theta)=\mathrm{Var}(F(X_1+\theta))=\int_{-\infty}^{\infty}F^2(x+\theta)f(x-\theta)dx$$
$$-(\int_{-\infty}^{\infty}F(x+\theta)f(x-\theta)dx)^2$$
$$=\int_{-\infty}^{\infty}F^2(x+2\theta)f(x)\,dx-(\int_{-\infty}^{\infty}F(x+2\theta)f(x)\,dx)^2 \tag{4·71}$$

又 U_n 的期望为

$$Eh(X_1,X_2)=P(X_1+X_2>0)=\int_{-\infty}^{\infty}F(x+2\theta)f(x)dx, \tag{4·72}$$

此式可通过先计算条件概率

$$P(X_1+X_2>0|X_1=x)=P(X_2>-x)=F(x+\theta),$$

然后用

$$P(X_1+X_2>0)=E\{P(X_1+X_2>0\mid X_1)\}=E(F(X_1+\theta))$$
$$=\int_{-\infty}^{\infty}F(x+\theta)f(x-\theta)d\theta$$
$$=\int_{-\infty}^{\infty}F(x+2\theta)f(x)dx$$

而得到。据（4·71），（4·72），再注意到此处相当于（3·18）式中的$m=2$，故若取

$$\mu(W^+,n,\theta,F)=\binom{n}{2}\int_{-\infty}^{\infty}F(x+2\theta)f(x)dx \qquad (4\cdot73)$$

及

$$\sigma(W^+,n,\theta,F)=\frac{2}{\sqrt{n}}\binom{n}{2}\left\{\int_{-\infty}^{\infty}F^2(x+2\theta)f(x)dx\right.$$
$$\left.-\left(\int_{-\infty}^{\infty}F(x+2\theta)f(x)dx\right)^2\right\}^{1/2}, \qquad (4\cdot74)$$

则将有

$$\left(\binom{n}{2}U_n-\mu(W^+,n,\theta,F)\right)/\sigma(W^+,n,\theta,F)\xrightarrow{\mathscr{L}}N(0,1).$$
$$(4\cdot75)$$

但$|B_n|\leqslant n$，而$\sigma(W^+,n,\theta,F)$为$n^{3/2}$的数量级。故有$B_n/\sigma(W^+,n,\theta,F)\to0$当$n\to\infty$。因此由（4·74）与（4·69）得

$$(W_n^+-\mu(W^+,n,\theta,F))/\sigma(W^+,n,\theta,F)\xrightarrow{\mathscr{L}}N(0,1).$$
$$(4\cdot76)$$

（4·76）说明，（4·73）和（4·74）的取法可行

有了以上的准备，就不难计算所考虑的三个检验的效率因子。对 t 检验有

$$\mu'(t,n,\theta,F)=\sqrt{n}/\delta,\ \sigma(t,n,\theta,F)=1.$$

于是按（4·61）得

$$K_t^2(F)=1/\delta^2, \qquad (4\cdot77)$$

对符号检验B，有$\mu'(B,n,\theta,F)=nf(\theta)$，$\sigma(B,n,\theta,F)=\frac{\sqrt{n}}{2}$。

于是

$$K_B^2(F)=4f^2(0). \tag{4·78}$$

对W^+检验，有

$$\mu'(W^+,n,\theta,F)=2\binom{n}{2}\int_{-\infty}^{\infty}f(x+2\theta)f(x)dx\to 2\binom{n}{2}\left(\int_{-\infty}^{\infty}f^2(x)dx\right),$$

当$\theta\to 0$。又

$$\lim_{\theta\to 0}\sigma(W^+,n,\theta,F)=\frac{2}{\sqrt{n}}\binom{n}{2}\Big(\int_{-\infty}^{\infty}F^2(x)dF(x)$$
$$-\Big(\int_{-\infty}^{\infty}F(x)dF(x)\Big)^2\Big)^{1/2}$$
$$=\frac{2}{\sqrt{n}}\binom{n}{2}\Big(\int_0^1 t^2dt-\Big(\int_0^1 tdt\Big)^2\Big)^{1/2}=\frac{1}{\sqrt{3n}}\binom{n}{2},$$

由此得出

$$K_{W^+}^2(F)=12\Big[\int_{-\infty}^{\infty}f^2(x)dx\Big)^2. \tag{4·79}$$

由（4·77）—（4·79)，用公式（4·65)，即可得到t，B，W^+这三个检验之间的ARE。现就几个重要的分布F列出其数值如下：

分布 F	密　　度	ARE(W^+,t,F)	ARE(B,t,F)
正态	$e^{-x^2/2}/\sqrt{2\pi}$	$3/\pi$	$2/\pi$
均匀R(-1，1)	$\frac{1}{2}I(-1<x<1)$	1	1/3
Logistic	$e^{-x}(1+e^{-x})^2$	$\pi^2/9$	$\pi^2/12$
重指数	$\frac{1}{2}e^{-\lvert x\rvert}$	3/2	2

通过对这几个典型分布的计算看出，Wilcoxon 符号秩和检验与传统的t检验比，有相当的优势：在这里计算的几个值中，有三个大于或等于1。只有在正态分布（t检验是专门针对这一场合

的）下，此值略小于1但很接近1．因此，W^+检验由于其在原假设下的分布无关性，保证了它不致因模型偏离正态而犯大错误，其代价只是在模型真为正态时，效率略为降低。即使表面上看来很粗糙的符号检验，其与 t 检验的对比也是有好有坏。这些结果使我们对非参数方法的性能具有信心。下面的结果更增强了这一点：对任何具有密度 f 与有限方差的对称分布 F，总有

$$\mathrm{ARE}(W^+,t,F)\geqslant 0.864. \tag{4·80}$$

换句话说，使用W^+与 t 对比，在效率上的损失不会超过13.6%。为证(4·86)，不妨假定F的方差为1．因若F的方差为δ，则$F(\delta x)$的方差为1．对分布 $F(\delta x)$ 而言，$K_t^2(F(\delta x))=\delta^2K_t^2(F(x))$，$K_{W^+}^2(F(\delta x))=\delta^2K_{W^+}^2(F(x))$。故 ARE $(W^+,t,F(\delta x))=$ARE $(W^+t,F(x))$．因此我们可用$F(\delta x)$代替 $F(x)$去讨论，而转化为方差1的情况。

根据（4·77）和（4·79），为证（4·80），要在

$$f(x)\geqslant 0,\quad f(-x)=f(x),\quad \int_{-\infty}^{\infty}f(x)dx=1,\quad \int_{-\infty}^{\infty}x^2f(x)dx=1 \tag{4·81}$$

的条件下，去证明

$$12\left(\int_{-\infty}^{\infty}f^2(x)dx\right)^2\geqslant 0.864 \tag{4·82}$$

为此，取

$$f_0(x)=\frac{3}{20\sqrt{5}}(5-x^2)I(|x|<\sqrt{5}), \tag{4·83}$$

以及

$$f_1(x)=\frac{3}{20\sqrt{5}}(5-x^2),-\infty<x<\infty,$$

有

$$\int_{-\infty}^{\infty}f^2(x)dx=\int_{-\infty}^{\infty}f_0^2(x)+\int_{-\infty}^{\infty}(f(x)-f_0(x))^2dx$$
$$+2\int_{-\infty}^{\infty}f_0(x)(f(x)-f_0(x))dx. \tag{4·84}$$

由（4·81）知

$$\int_{-\infty}^{\infty} f_1(x)f(x)dx = \frac{3}{20\sqrt{5}}(5-1) = \frac{3}{5\sqrt{5}}$$

$$= \int_{-\infty}^{\infty} f_0^2(x)dx, \qquad (4\cdot85)$$

由于$f_0(x)-f_1(x)$及$f(x)$在全直线上非负，有

$$\int_{-\infty}^{\infty}(f_0(x)-f_1(x))f(x)dx \geqslant 0, \qquad (4\cdot86)$$

由(4·85)及(4·86)推出 $\int_{-\infty}^{\infty} f_0(x)(f(x)-f_0(x))dx \geqslant 0$，这与(4·84)结合，即知

$$\int_{-\infty}^{\infty} f^2(x)dx \geqslant \int_{-\infty}^{\infty} f_0^2(x)dx = \frac{3}{5\sqrt{5}},$$

因此 $12(\int_{-\infty}^{\infty} f^2(x)dx^2) \geqslant 12\left(\frac{3}{5\sqrt{5}}\right)^2 = \frac{108}{125} = 0.864$。

故（4·82）对任何满足（4·81）的f都成立，从而证明了(4·80)。

例4·7 设$X_1,\cdots,X_{n_1}$和$Y_1,\cdots,Y_{n2}$分别是抽自分布$F(x)$与$F(x-\theta)$的简单样本，为检验假设$\theta\leqslant 0$，考虑两个检验法：其一是通常的两样本t检验，其统计量为

$$T_n = \sqrt{\frac{n_1 n_2}{n}}(\overline{Y}_{n_2}-\overline{X}_{n_1})/s_{n_1n_2}$$

此处$n=n_1+n_2$，$\overline{Y}_{n2}$和$\overline{X}_{n_1}$分别是Y样本与X样本的样本均值，而$s_{n_1n_2}^2 = \frac{1}{n-2}\left(\sum_{i=1}^{n_1}(X_i-\overline{X}_{n_1})^2 + \sum_{i=1}^{n_2}(Y_i-\overline{Y}_{n_2})^2\right)$。另一个是Wilcoxon秩和检验$W$，其统计量为

$$W_n = R_1 + \cdots + R_{n_2}$$

此处R_i为Y_i在合样本中的秩。以下我们假定分布F有密度f，且F有有限的方差δ^2。

与一样本情况类似，对t检验可取

$$\mu(t,n,\theta,F) = \sqrt{\frac{n_1n_2}{n}}\theta/\delta,\ \sigma(t,n,\theta,F)=1, \qquad (4\cdot87)$$

它就能满足条件（4·53）及（4·55）。

对W检验而言，通过例3·3，可知（见(3·10)式）

$$W_n = \frac{1}{2}n_2(n_2+1)+n_1n_2U_{n_1n_2},$$

其中$U_{n_1n_2}$是以（3·8）为核的U统计量.就这一部分而言，可用定理3·2.与例4·6类似的算法，算出

$$\sigma_{10}^2 = \int_{-\infty}^{\infty}(1-F(x-\theta))^2dF(x) - \left(\int_{-\infty}^{\infty}(1-F(x-\theta))dF(x)\right)^2,$$

$$\sigma_{01}^2 = \int_{-\infty}^{\infty}F^2(x+\theta)dF(x) - \left(\int_{-\infty}^{\infty}F(x+\theta)dF(x)\right)^2.$$

又 $EU_{n_1n_2} = \int_{-\infty}^{\infty} F(x+\theta)dF(x)$.因此，由定理3·2知，若取

$$\mu(W,n,F,\theta)=\frac{1}{2}n_2(n_2+1)+n_1n_2\int_{-\infty}^{\infty}F(x+\theta)dF(x), \tag{4·88}$$

$$\sigma(W,n,F,\theta)=\sqrt{n_1n_2n/12}, \tag{4·89}$$

且设

$$\lim_{n\to\infty} n_1/n=\lambda,\quad 0<\lambda<1, \tag{4·90}$$

则知当$\theta_n\to 0$时，有

$$[W_n-\mu(W,n,F,\theta_n)]/\sigma(W,n,F,\theta) \xrightarrow{\mathscr{L}} N(0,1).$$

这说明对W检验，（4·88）和（4·89）的选择正确.据（4·87)—（4·89)，且在（4·90）的假定下，就不难算出这两个检验的效率因子分别为

$$K_t^2(F)=\lambda(1-\lambda)/\delta^2, \tag{4·91}$$

$$K_W^2(F)=12\lambda(1-\lambda)\left(\int_{-\infty}^{\infty}f^2(x)dx\right)^2. \tag{4·92}$$

由（4·91）和（4·92),得

$$\mathrm{ARE}(W,t;\ F)=12\delta^2\left(\int_{-\infty}^{\infty}f^2(x)dx\right)^2. \tag{4·93}$$

值得注意的是此值与λ无关，且等于ARE(W^+,t; F)（当然后一量要求F关于0对称）.理论上可以证明，这不是一个巧合。既然ARE(W,t; F)等于ARE($W^+,t;F$)，在前例中关于W^+和t的对比情况所说的一切，可一字不改地移到此处。特别，(4·93)右边的最小值为0.864。

注意在以上两例中，在验证定理4·11的条件时，有些细节被忽略了。比方说，在例4·7中为计算$\mu'(W,n,\theta,F)$，我们是把$\int_{-\infty}^{\infty}F(x+\theta)dF(x)$在积分号下对$\theta$求导得$\int_{-\infty}^{\infty}F'(x+\theta)dF(x)$

$=\int_{-\infty}^{\infty}f(x+\theta)f(x)dx$，而这需要细致的分析论证。如果我们不计较这些细节，那么关于一般的两样本秩检验也不难算出其效率因子，结果如下：设检验L在样本大小为n(实际上，$n=n_1+n_2$，为合样本大小）时的统计量选为$L_n=\sum_{i=1}^{n_2}a_n(R_i)$，其中$a_n(i)=\varphi\left(\frac{i}{n+1}\right)$且$\varphi\in SS$（见定理4·4前一段的说明），则将有

$$K_L^2(F)=\lambda(1-\lambda)\left(\int_{-\infty}^{\infty}\varphi'(F(x))f^2(x)dx\right)^2\Big/\left(\int_0^1\varphi^2(x)dx-\left(\int_0^1\varphi(x)dx\right)^2\right). \qquad (4\cdot94)$$

这里设$\lim_{n\to\infty}n_1/n$存在且等于λ，又F有密度f。由表达式(4·94)已见，我们假定了φ'存在。这公式从形式推导上说易于从定理4·7推出（建议读者自己作一下），但涉及定理4·11中条件的仔细验证，则有不少繁琐的工作要做。

还可以证明：若计分函数用$a_n(i)=E\varphi(U_{ni})$去定义，其中$U_{n1}\leqslant\cdots\leqslant U_{nn}$是从(0,1)均匀分布中抽得的次序样本，则由之所确定的检验的效率因子也是(4·94)。

例4·8 对Fisher-Yates检验和Van der waerden检验，有，$=\Phi^{-1}$，Φ^{-1}是$N(0,1)$的分布Φ的反函数。记$g(x)=e^{-x^2/2}/\sqrt{2\pi\varphi}$

则有$\varphi'(u)=(g(\Phi^{-1}(u)))^{-1}$.又

$$\int_0^1\varphi(u)du=\int_{-\infty}^{\infty}xg(x)dx=0,$$

$$\int_0^1\varphi^2(u)du=\int_{-\infty}^{\infty}x^2g(x)dx=1,$$

于是由（4·94）得到

$$K_V^2(F)=K_{FY}^2(F)=\lambda(1-\lambda)\left(\int_{-\infty}^{\infty}\frac{f^2(x)}{g(\Phi^{-1}(F(x)))}dx\right)^2 \tag{4·95}$$

特别，当F为正态分布$N(a,\delta^2)$时，算出

$$K_V^2(F)=K_{FY}^2(F)=\lambda(1-\lambda)/\delta^2. \tag{4·96}$$

由此式与（4·91），即得

$$\text{ARE}(V,t,\text{正态})=\text{ARE}(FY,t,\text{正态})=1. \tag{4·97}$$

即从大样本角度看，Fisher-Yates和Van der Waerden检验与t检验在总体为正态时，有相同的效率，但后者不具备“分布无关”的优点，因而可以说，在样本大小较大时，用FY和V检验比用t检验更合理，更进一步，可以证明在F的方差为1的限制下

$$\inf_F\int_{-\infty}^{\infty}\frac{f^2(x)}{g(\Phi^{-1}(F(x)))}dx=1, \tag{4·98}$$

且最小值在F为正态时达到.所以，从大样本观点看，FY和V检验在任何情况下都不劣于t检验。

公式（4·94）可用于解决下述有实际意义的问题：针对一特定的分布F（其密度f存在），找一个秩检验，其在分布F处的效率因子比任何其他秩检验在F处的效率因子都大。据（4·94）可以证明：它就是以$a_n(i)=\varphi_F\left(\frac{i}{n+1}\right)$（或$a_n(i)=E(\varphi_F(U_{ni}))$也可以）为计分函数所确定的秩检验，此处

$$\varphi_F(u)=-f'(F^{-1}(u))/f(F^{-1}(u)). \tag{4·99}$$

有了这个公式，我们可按照所设想的最可能的总体分布F去选择效率因子最高的秩检验。这公式的成立当然有一些条件，至少从

（4·99）看到，这些条件包括：F有密度f，且f的导数存在，又分布F严格增加（这等于要求f在（$-\infty$，∞）处处大于0），因而F^{-1}存在。例如，当F为正态时，由（4·99）可算出$\varphi_F(u)=\Phi^{-1}(u)$，这相应于Fisher-Yates或Van der Waerden检验。

关于对称中心的检验问题，也可得到类似的一般结果，其严格理论比两样本问题还要复杂，这里只引述其结果。

设我们用线性符号秩统计量$L_n^+=\sum_{i=1}^{n}\psi_i\varphi\left(\frac{R_i^+}{n+1}\right)$去检验总体分布$F(x-\theta)$中的$\theta=0$，此处$F(x)$关于0对称。定义

$$q_F(u)=f'\left(F^{-1}\left(\frac{1+u}{2}\right)\right)/f\left(F^{-1}\left(\frac{1+u}{2}\right)\right),\quad 0<u<1$$

此处$f=F'$。则这个检验L^+的效率因子为

$$K_{L^+}^2(F)=\left(\int_0^1 q_F(u)\varphi(u)du\right)^2\Big/\int_0^1\varphi^2(u)du.\quad(4\cdot100)$$

对Wilcoxon符号秩和检验，有$\varphi(u)=(n+1)u$。读者不难据（4·100）算出其效率因子如（4·79）。

另外，在例4·7中我们曾指出，Wilcoxon一、二样本检验的效率因子只相差一个常数因子$\lambda(1-\lambda)$，而这不是偶然的巧合，事实上，在一定的条件下可以证明下面的结果：若关于同一分布F的一、二样本秩检验L和L^+都由同一个φ决定，且在两样本问题中有$n_1/n\to\lambda$，则

$$K_L^2(F)=\lambda(1-\lambda)K_{L^+}^2(F).$$

三、局部最优秩检验

上一段所讨论的准则——渐近相对效率，是大样本性质的。本段将引进另一个判断检验的优良性的准则——局部最优性。它是小样本性质的。

设我们有了样本$Z_1,\cdots,Z_n$而要检验某个原假设H。当H成立时总体分布属于一定的分布族$\mathscr{F}$。对立假设一般也是一个很大的分布族，我们从其中挑出一个可由一实参数θ去刻划的子族（如

在两样本问题中，当原假设不成立时总体分布可表为(F,G)，其中$F\neq G$.考虑子族$\{(F(x), F(x-\theta)):\theta\geqslant 0\}$，其中$F$已知.当$\theta=0$时属于原假设，而$\theta>0$时属对立假设）. 设$\theta=\theta_0$时属原假设，而$\theta>\theta_0$时属对立假设. 我们希望找到一个水平$\alpha$的秩检验，其在$\{\theta>\theta_0\}$这部分对立假设上一致最优.这种检验一般不存在. 于是我们退而求其次：找这样一个水平α的秩检验，使它对某个$\varepsilon>0$，在$\{\theta_0<\theta<\theta_0+\varepsilon\}$这个局部上达到最优.如这种秩检验存在，则它可称为"局部一致最优"的. 可惜的是，即使这种检验也往往不存在，如是我们再退一步：不要求ε固定，而只要求在一个"无限小"的区间内达到最优. 这个考虑引导到下述局部最优秩检验的定义.

定义4·4 以$\mathscr{A}_\alpha$记原假设H的所有真实水平α秩检验之集. 对任何$S\in\mathscr{A}_\alpha$，以$\beta_S(\theta)(\theta\geqslant\theta_0)$为其在对立假设$\{\theta<\theta_0\}$上的功效函数.设$S_0\in\mathscr{A}_\alpha$.若对任何$S\in\mathscr{A}_\alpha$都有$\beta'_S(\theta_0)\leqslant\beta'_{S_0}(\theta_0)$，则称$S_0$是$H$的针对所述对立假设的水平$\alpha$的局部最优秩检验（Local Most Powerful Rank Test，简记为LMPRT）.

这个定义的含义不难从上面的说明得到理解.先设想$\beta'_S(\theta_0)<\beta'_{S_0}(\theta_0)$.由于$\alpha$是真实水平，有$\beta_S(\theta_0)=\beta_{S_0}(\theta_0)=\alpha$（注意$\theta=\theta_0$属于原假设）.故由$\beta'_S(\theta_0)<\beta'_{S_0}(\theta_0)$知，当$\theta>\theta_0$但$\theta-\theta_0$充分小时有$\beta_S(\theta)<\beta_{S_0}(\theta)$.这正是上述关于局部最优的要求.若$\beta'_S(\theta_0)=\beta'_{S_0}(\theta_0)$，则$\beta_{S_0}(\theta)$可能小于$\beta_S(\theta)$，差$\beta_S(\theta)-\beta_{S_0}(\theta)$也只能是$\theta-\theta_0$的高级无穷小量，即在$\theta_0$的很小的邻域内当比较功效的线性主部时，$S$仍不优于$S_0$.

下面我们来求两样本问题的LMPRT，对立假设子族选为$\{(F(x), F(x-\theta)):\theta>0\}$.有以下的结果:

定理4·12 以$\xi_1\leqslant\cdots\leqslant\xi_n$记抽自分布$F(x)$的大小为$n$的次序样本，则相对于上述对立假设子族而言，LMPRT是由下述计分函数确定的秩检验:

$$a_n(i)=-E(f'(\xi_i)/f(\xi_i)),\quad i=1,\cdots,n, \qquad (4\cdot 101)$$

此处 $f=F'$.

为证本定理，需要下面的引理。

引理4·1 设 $X_1,\cdots,X_{n_1}$ 和 $Y_1,\cdots,Y_{n_2}$ 分别是从密度 f 和 g 中抽出的简单样本，H 为一分布，其密度 h 处处大于0。记 $(X_1,\cdots,X_{n_1},Y_1,\cdots,Y_{n_2})=(Z_1,\cdots,Z_n)$。以 R_i 记 Z_i 在 $(Z_1,\cdots,Z_n)$ 中的秩，$R=(R_1,\cdots,R_n)$。则对 $(1,2,\cdots,n)$ 的任一置换 $r=(r_1,\cdots,r_n)$ 有

$$P(R=r)=\frac{1}{n!}E\left\{\prod_{i=1}^{n_1} f(V_{r_i})\prod_{i=n_1+1}^{n} g(V_{r_i})\Big/\prod_{i=1}^{n} h(V_i)\right\},\tag{4·102}$$

此处 $V_1\leqslant\cdots\leqslant V_n$ 是从密度 h 中抽出的、大小为 n 的次序样本。

证明　以 B 记 n 维欧氏空间的子集

$B=\{(v_1,\cdots,v_n):v_1<\cdots<v_n\}$,

则有

$$P(R=r)=\int_B\prod_{i=1}^{n_1} f(v_{r_i})\prod_{i=n_1+1}^{n} g(v_{r_i})dv_1\cdots dv_n$$

$$=\frac{1}{n!}\int_B\frac{\prod_{i=1}^{n_1} f(v_{r_i})\prod_{i=n_1+1}^{n} g(v_{r_i})}{\prod_{i=1}^{n} h(v_i)}\left(n!\prod_{i=1}^{n} h(v_i)\right)dv_1\cdots dv_n\tag{4·103}$$

但据(2·11)，$(V_1,\cdots,V_n)$ 的密度，在 B 上为 $n!\prod_{i=1}^{n} h(v_i)$，而在 B 外则为0。故上式可写为(4·102)。证毕。

现转到定理4·12的证明。任何一个秩检验，等价于 $(1,2,\cdots,n)$ 的一切可能的置换（共 $n!$ 个）之集的一个子集 J。意思是，当且仅当秩统计量 R 的取值 r 落在 J 内时，才否定原假设。把这个秩检验也记为 J。按引理4·1检验 J 的功效为

$$\beta_J(\theta)=\sum_{r\in J}\frac{1}{n!}E\left(\prod_{i=n_1+1}^{n}\frac{f(\zeta_{r_i}-\theta)}{f(\zeta_{r_i})}\right)\tag{4·104}$$

此式是由在（4·103）中取$g(x)=f(x-\theta)$及$h=f$（这就要求f处处大于0）得来的。易见

$$\beta'_J(\theta_0)=\frac{1}{n!}\sum_{r\in J}\sum_{i=n_1+1}^{n}E(-f'(\xi_{r_i})/f(\xi_{r_i}))。\quad（4·105）$$

由（4·105）看出：为使$\beta'_J(\theta_0)$最大，应把那些使表达式

$$\sum_{i=n_1+1}^{n}E(-f'(\xi_{r_i})/f(\xi_{r_i}))$$

尽可能大的置换r收到否定域J中去.这就证明了本定理。

如用（0,1）均匀分布的次序样本$U_{n1}\leqslant\cdots\leqslant U_{nn}$，可将（4·101）写为

$$a_n(i)=-E(f'(F^{-1}(U_{ni}))/f(F^{-1}(U_{ni})))。$$

此与（4·99）对照可知：针对F为LMPRT的秩检验，也是在F处有最大效率因子的秩检验。初一看这似属巧合，实则不然。因为归根到底，二者都是基于在原假设点θ_0近旁处功效值，大者为优。对ARE这个准则而言，定理4·12规定的秩检验，与由计分函数$a_n(i)=-f'(F^{-1}(\frac{i}{n+1}))/f(F^{-1}(\frac{i}{n+1}))$决定的秩检验无高低之分别。在此则不然：局部最优秩检验只有一个，即定理4·12所决定者，其他都不是。这样，在F为正态分布时，按"局部最优"这个准则，你可以说Fisher-Yates检验优于Van der Waerden检验。

对一样本问题，也有类似的结果，但论证更为复杂，此处不细述了。

§4·3 多样本问题与随机区组秩检验

本节的内容是讲述秩方法在简单的方差分析问题中的应用。

一、多样本问题

多样本问题是两样本问题的直接推广：设有m个一维总体。

其分布分别记为$F_1,\cdots,F_m$.从第i个总体中抽出简单样本$X_{i1},\cdots X_{in_i}$，$i=1,\cdots,m$，又假定这$n=n_1+\cdots+n_m$个样本全体独立。要依据这些样本去检验假设

$$H:F_1=F_2=\cdots=F_m, \tag{4·106}$$

从方差分析的观点看，这m个总体可看成是一个因素的m个水平。假设（4·106）的意义是这m个水平无差别，或者说，该因素无效应。

1. 一般对立假设的情况.就是说，对立假设无方向性。当H不成立时，只简单地知道$F_1,\cdots,F_m$并非全恒等，其他一无所知。

用秩统计量检验（4·106）的方法，可以由两样本情况得到启发。以R_{ij}记X_{ij}在合样本$\{X_{ij}:j=1,\cdots,n_i,i=1,\cdots,m\}$中的秩，目前我们暂假定分布$F_1,\cdots,F_m$都处处连续，因而不发生结的问题。给定计分函数$a_n(\cdot)$，而令

$$L_{ni}=\sum_{j=1}^{n_i}a_n(R_{ij}),\quad S_{ni}=(L_{ni}-\mu_{ni})/\sigma_{ni},\quad i=1,\cdots,m, \tag{4·107}$$

此处μ_{ni}和σ_{ni}^2分别是L_{ni}在H成立时的数学期望与方差。按公式（4·3）和（4·4），有

$$\mu_{ni}=n_i\bar{a}_n,\qquad \sigma_{ni}^2=\frac{1}{n-1}\frac{n_in_i'}{n}D_n, \tag{4·108}$$

此处$\bar{a}_n=\sum_{i=1}^{n}a_n(i)/n$，$n_i'=n-n_i$，$D_n=\sum_{i=1}^{n}(a_n(i)-\bar{a}_n)^2$。

按（4·108），若H成立，则L_{ni}/n_i，$i=1,\cdots,m$，都有相同的期望$\bar{a}_n$，因此$(L_{ni}/n_i-\bar{a}_n)^2$应倾向于小.故把这些表达式作加权和（按各样本大小$n_1,\cdots,n_m$加权），并加以规则化，得统计量

$$T_n=\frac{n-1}{D_n}\sum_{i=1}^{m}n_i(L_{ni}/n_i-\bar{a}_n)^2。 \tag{4·109}$$

它作为衡量样本$\{X_{ij}\}$与假设H的偏离程度的一种指标，T_n愈大，偏离愈显著。因此一个合理的检验是：当

$$T_n>C \tag{4·110}$$

时否定原假设H。C根据T_n在H下的分布，及给定的α定出。

当$n_1,\cdots,n_m$都较小时，直接求T_n的精确分布（在H成立下）尚属可行。如n_i较大，则只好诉诸极限分布。往下我们来证明：如果$\{(a_n(1),\cdots,a_n(n)):n=1,2,\cdots\}$满足定理4·4或4·5中的条件，而且当$n\to\infty$时

$$n_i/n\longrightarrow\rho_i>0\text{存在,}\quad i=1,\cdots,m, \tag{4·111}$$

则当$n\to\infty$时，在H成立之下，有

$$T_n\xrightarrow{\mathscr{L}}\chi^2_{m-1}. \tag{4·112}$$

这里χ^2_{m-1}是自由度$m-1$的中心χ^2分布。

为证此要用到定理4·6.据该定理（在H成立时，下同），在此处所设条件下有

$$S_n\equiv((L_{n1}-\mu_{n1})/\sigma_{n1},\cdots,(L_{n,m-1}-\mu_{n,m-1})/\sigma_{n,m-1})$$
$$\xrightarrow{\mathscr{L}}N(0,\Lambda) \tag{4·113}$$

此处$\Lambda=(\lambda_{ij})$为$m-1$阶方阵，λ_{ij}据定理4·6中的表达式（注意此处$(c^{(k)}_{n1},\cdots,c^{(k)}_{nn})$是$(0,\cdots,0;\cdots;0,\cdots,0;1,\cdots,1;0,\cdots0;\cdots;0,\cdots,0)$：一共$m$段，只第$k$段有$n_k$个1，其余全为0)及(4·111)易算得为

$$\lambda_{ii}=1,\ \lambda_{ij}=-(\rho_i\rho_j/((1-\rho_i)(1-\rho_j)))^{1/2},\quad i,\quad j=1,\cdots,m. \tag{4·114}$$

现在要利用概率论中的一个定理：设ξ服从n维正态分布$N(0,\Lambda)$，其中矩阵Λ非异，则$\xi'\Lambda^{-1}\xi$服从自由度为n的χ^2分布χ^2_n.由这个定理及（4·113），得知当$n\to\infty$时有

$$S'_n\Lambda^{-1}S_n\xrightarrow{\mathscr{L}}\chi^2_{m-1} \tag{4·115}$$

直接计算易证明：

$$\Lambda^{-1}=\begin{pmatrix} 1-\rho_1 & & & 0 \\ & 1-\rho_2 & & \\ 0 & & \cdots & \\ & & & 1-\rho_{m-1} \end{pmatrix}+dd'/\rho_m,$$

其中$d=(\sqrt{\rho_1(1-\rho_1)},\cdots,\sqrt{\rho_{m-1}(1-\rho_{m-1})})'$. 以这个$\Lambda^{-1}$代入(4·115)的左边，计算其表达式，但是把$\Lambda^{-1}$中的$\rho_i$改成$n_i/n$。由于$n_i/n\longrightarrow\rho_i$，这一修改将不影响(4·115)的成立。得到的表达式是(据(4·107)，(4·108))：

$$\sum_{i=1}^{m-1}\frac{n-1}{D_n}n_i(L_{ni}/n_i-\bar{a}_n)^2+\frac{n-1}{n_mD_n}\sum_{i=1}^{m-1}\left(L_{ni}-n_i\bar{a}_n\right)^2. \quad (4·116)$$

从L_{ni}的定义(见(4·107))可知$\sum_{i=1}^{m}L_{ni}=\sum_{i=1}^{n}a_n(i)=n\bar{a}_n$，因此

$\sum_{i=1}^{m-1}(L_{ni}-n_i\bar{a}_n)=-(L_{nm}-n_m\bar{a}_n)$。因此(4·116)就是(4·109)所定义的$T_n$。于是证明了(4·112)。

根据这个结果，当n_i较大时，(4·110)中的C可近似地取为$\chi^2_{m-1}(\alpha)$，此处α为给定的检验水平，而$\chi^2_{m-1}(\alpha)$为χ^2_{m-1}的$100(1-\alpha)\%$分位点。

例 4·9 检验(4·110)的一个重要特例是$a_n(i)=i$。所产生的检验叫Kruskal-Wallis(KW)检验，是他们两人在1952年提出的。这相当于两样本中的Wilcoxon检验，其统计量为

$$T_n=\frac{12}{n(n+1)}\sum_{i=1}^{m}n_i(R_i-\frac{n+1}{2})^2,\ R_i=\sum_{j=1}^{n_i}R_{ij}/n_i, \quad (4·117)$$

R_i即第i组样本$X_{i1},\cdots,X_{in_i}$在合样本中之秩之平均。在多样本问题中，与t检验相当的检验是F检验。它基于所谓F统计量

$$\mathscr{F}_n=\frac{1}{m-1}\sum_{i=1}^{m}n_i(\bar{X}_i-\bar{X})^2/\frac{1}{n-m}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_i)^2, \quad (4·118)$$

这里$\bar{X}_i=\sum_{j=1}^{n_i}X_{ij}/n_i$，是第 i 组样本的平均值．在$F_1,\cdots,F_m$为相同的正态分布时，$\mathscr{F}_n$ 服从自由度为 $m-1$ 及 $n-m$ 的F 分布$F_{m-1,n-m}$，因此基于此统计量的、水平 α 的检验的否定域是

$$\mathscr{F}_n>F_{m-1,n-m}(\alpha).$$

从渐近相对效率的一般定义出发，可算出在位置参数型的对立假设下（即形如 $F_i(x)=F(x-\theta_i)$，$i=1,\cdots,m$的对立假设，其中$\theta_1,\cdots,\theta_m$ 不全相同），由（4·110）确定的秩检验T对F检验的ARE．有趣的是，计算结果与两样本情况完全相同，例如

$$\mathrm{ARE}(KW,\ \mathscr{F};\ F)=12\sigma^2\left(\int_{-\infty}^{\infty}f^2(x)dx\right)^2$$

此处 $f=F'$，而 σ^2 为分布F的方差．因此，在§4·2中讲到过的有关秩检验与 t 检验的对比的一切，可一字不改地移于此处．

2．结存在的情况．当结存在时，要对由公式（4·109）定义的 T_n 作些修改（如用随机法定秩，不须作任何修改．此处讨论的是平均法）．修改步骤如下：

a．把样本$\{X_{ij}\}$排成一列：$X_{11},\cdots,X_{1n_1},\cdots,X_{m1},\cdots,X_{mn_m}$并以单足标记之：$Z_1,\cdots,Z_n$．集合$\{1,2,\cdots,n\}$分解为$q$个互不相交的子集 $J_1,\cdots,J_q$，使当且仅当u和v都落在同一个 J_r 之内时，才有 $Z_u=Z_v$．又若$i\in J_r$，$j\in J_s$而$r<s$，则$Z_i<Z_j$．J_r中所含元素的个数记为 τ_r．$(\tau_1,\cdots,\tau_q)$ 就是在§4·1的三段中提到过的结统计量．

b．从$a_n(\cdot)$出发定义 $\bar{a}_n(\cdot)$如下：

$$\bar{a}_n(i)=\sum_{j=\tau_1+\cdots+\tau_{r-1}+1}^{\tau_1+\cdots+\tau_r}a_n(j)/\tau_r,$$

$$\text{当}\tau_1+\cdots+\tau_{r-1}+1\leqslant i\leqslant\tau_1+\cdots+\tau_r$$

c．在（4·109）的 T_n 定义中，把各样本原来的秩改成经修正后的秩，而 $a_n(\cdot)$ 改为 $\bar{a}_n(\cdot)$（但D_n 仍维持不改，即仍为$\sum_{i=1}^{n}(a_n(i)-\bar{a}_n)^2$）．修改后算出的 T_n 暂记为$\bar{T}_n$．

d．计算

$$T_n^* = \left(\sum_{i=1}^{n} (a_n(i) - \bar{a}_n)^2 / \sum_{i=1}^{n} (\tilde{a}_n(i) - \bar{a}_n)^2 \right) T'_n,$$

则可证：若记分函数 $a_n(\cdot)$ 满足定理4·4或4·5的条件，且（4·111）成立，则在原假设之下仍有．$T'_n \xrightarrow{\mathscr{L}} \chi^2_{m-1}$，因而以 $\{T_n^* > \chi^2_{m-1}(\alpha)\}$ 为否定域的检验具有渐近水平 α。

我们留给读者证明：对 Kruskal-Wallis 检验，即 $a_n(i) = i$，有

$$\sum_{i=1}^{n} (\tilde{a}_n(i) - \bar{a}_n)^2 = \frac{n(n^2-1)}{12} - \frac{1}{12} \sum_{r=1}^{q} (\tau_r^3 - \tau_r), \qquad (4\cdot119)$$

因此，修正后的 Kruskal-Wallis 统计量为

$$T_n^* = \{1 - \sum_{r=1}^{q} (\tau_r^3 - \tau_r)/(n^3 - n)\}^{-1} \frac{12}{n(n+1)} \sum_{i=1}^{m} n_i \left(R_i^* - \frac{N+1}{2}\right)^2, \qquad (4\cdot120)$$

这里 R_i^+ 是 $X_{i1}, \cdots, X_{in_i}$ 在合样本中用平均法修正后的秩的和，再除以 n_i。

下面再通过一个数字例以解释上述步骤．

例4·10 从三个总体中分别抽出大小为 5,5,7 的样本，结果为

5,4,6,4,6；　7,3,5,6,5；　2,3,3,1,2,1,1．

(1) 把合样本按由小到大排列为

1,1,1,2,2,3,3,3,4,4,5,5,5,6,6,6,7，

有 $q = 7$（长为1的结也算上），$(\tau_1, \cdots, \tau_7) = (3,2,3,2,3,3,1)$．

(2) 此处 $a_n(i) = i$．调整后各样本的秩分别是：三个1都是2，两个2都是4.5，其余3,4,5,6,7分别为7、9.5、12、15和17．由此算出，（4·120）中的 R_i^* 为：$R_1^* = (12+9.5+15+9.5+15)/5 = 61/5$，$R_2^* = 63/5$，$R_3^* = 29/7$．

(3) 以上述诸 τ_r 及 R_i^* 之值，以及 $n_1 = n_2 = 5$，$n_3 = 7$ 和

$n=17$代入（4·120），得

$$T_n^* = \left(1-\frac{108}{4896}\right)^{-1}\frac{12}{306}\left(5\times\left(\frac{61}{5}-9\right)^2+5\times\left(\frac{63}{5}-9\right)^2+7\times\left(\frac{29}{7}-9\right)^2\right)=10.7010.$$

此处 $m-1=2$。查 χ^2 分布表，得 $\chi_2^2(0.01)=9.210$。故即使在 $\alpha=0.01$ 的水平上也要否定各分布相同的原假设。

本例的最后公式（4·120）比较简单，不必依次经历上述一般的步骤 a、b、c、d。

3. 对立假设有序的情况

在有些情况下，根据问题的实际背景，有理由认为：当原假设（即各分布相同）不成立时，各总体变量的取值有沿一方向增长的趋势．就是说，若以 $X_1,\cdots,X_m$ 记这m个总体随机变量，则当原假设不成立时，有

$$X_u \overset{r}{>} X_v，当\ u>v \tag{4·121}$$

见定义3·3。

前面构造的秩检验（4·110）当然也可用于对付这种对立假设，但检验（4·110）是针对"一切"对立假设的，没有用到（4·121）的特殊性，利用这种特殊性，可构造出更富针对性的检验．方法如下：任取 $i<j$，考虑两组样本

$$X_{i1},\cdots,X_{in_i};\ X_{j1},\cdots,X_{jn_j}, \tag{4·122}$$

把 $X_{j1},\cdots,X_{jn_j}$在合样本（4·122）中的秩的和记为$R_n(i,j)$。因为在对立假设下当 $i<j$ 时$X_j \overset{r}{>} X_i$，故这时 $R_n(i,j)$ 会倾向于更大（相对于原假设下）．令

$$R(n)=\sum_{1\le i<j\le m} R_n(i,j), \tag{4·123}$$

此处仍以 n 记 $n_1+\cdots+n_m$．据以上的分析，当对立假设（4·121）成立时，$R(n)$倾向于大．由此提出一个检验，它以

$$R(n)>C \tag{4·124}$$

为否定域。

为要用大样本方法定C，就要定出统计量$R(n)$在原假设下的极限分布。可以证明：若条件（4·111）成立，则当$n\to\infty$时，有

$$(R(n)-A_n)/B_n \xrightarrow{\mathscr{L}} N(0,1), \tag{4·125}$$

其中

$$A_n=\sum_{1\leq i<j\leq m}\frac{1}{2}n_j(n_i+n_j+1),$$

$$B_n^2=\frac{1}{72}\{n^2(2n+3)-\sum_{i=1}^{m}n_i^2(2n_i+3)\}, \tag{4·126}$$

据（4·125），在给定的水平α之下，若n_i都较大，（4·124）中的C近似地可取为

$$C=A_n+B_nu_\alpha. \tag{4·127}$$

极限定理（4·125）的证明其实不难，但此处不给出了（参看习题14）。

以上是在分布连续的前提下讨论的。如果有结存在，则要作一定的修正，步骤如下：

a. 在（4·122）中把长大于1的结都找出来，设各结（长大于1者）之长为$t_1,\cdots,t_q$。又在长为t_r的结内，包含了第i总体的l_r个样本，$r=1,\cdots,q$。把$R(i,j)$修正为$R^*(i,j)=\sum_{u=1}^{n_i}\sum_{v=1}^{n_j}I(X_{iu}<X_{jv})+\frac{1}{2}\sum_{r=1}^{q}l_r(t_r-l_r)$ $R(n)$修正为$R^*(n)=\sum_{1\leq i<j\leq m}R^*(i,j)$（在此顺便指出：此处所定义的$R^*(i,j)$，就是在结存在的情况下，两组样本（4·122）的 Mann-Whitney 统计量。它对结不存在时的 Mann-Whitney 统计量（见（3·10）式及其下文）的修改，就是当某个X_{iu}与某个X_{jv}相等时，给以记分1/2。它与在结存在时取平均秩的 Wilcoxon 秩和统计量$\tilde{R}(i,j)$的关系为

$$\tilde{R}(i,j)-R^*(i,j)=\sum_{i=1}^{q}(t_i-l_i)(T_i-L_i)-\frac{1}{2}\sum_{i=1}^{q}(t_i-l_i)^2+$$

$n_j/2$，此处 t_i, l_i 已给了定义，而 $T_i = t_i + \cdots + t_i$，$L_i = l_1 + \cdots + l_i$。但应注意，此处要把长为 1 的"结"也算进来，即 t_i 可以为 1。又请读者验证：在不存在结时（一切 t_i 为 1 时），上式右边为 $n_j(n_j+1)/2$，即回到了前面我们在（3·10）式中所看到的情况。

b. 再考察全部样本

$X_{11}, \cdots, X_{1n1}, \cdots, X_{m1}, \cdots, X_{mn_m}$

找出其结统计量，记为$(\tau_1, \cdots, \tau_q)$。把 B_n^2 修正为

$$B_n^{*2} = \frac{1}{72}\{n(n-1)(2n+5) - \sum_{i=1}^{m} n_i(n_i-1)(2n_i+5)$$

$$- \sum_{r=1}^{q} \tau_r(\tau_r-1)(2\tau_r+5) + (36n(n-1)(n-2))^{-1}$$

$$\sum_{i=1}^{m} n_i(n_i-1)(n_i-2) \sum_{r=1}^{q} \tau_r(\tau_r-1)(\tau_r-2))$$

$$+ (8n(n-1))^{-1} \sum_{i=1}^{n} n_i(n_i-1) \sum_{r=1}^{q} \tau_r(\tau_r-1)\},$$

A_n修正为$A_n^* = (n^2 - n_1^2 - \cdots - n_m^2)/4$。

c. 在条件（4·111）之下，当$n\to\infty$时有

$$(R^*(n) - A_n^*)/B_n^* \xrightarrow{\mathscr{L}} N(0,1) \qquad (4\cdot128)$$

而公式（4·127）要用公式

$$C = A_n^* + B_n^* u_\alpha \qquad (4\cdot129)$$

去代替

例4·11 再考虑例4·10的数据，但把总体排序改为：原来最后的改为第一，原来第一的改为第二。

a. 算出 $R^*(1,2)=35$，$R^*(1,3)=34$，$R^*(2,3)=14$。由此得 $R^*(n)=35+34+14=83$。

b. 合样本结统计量为（3,2,3,2,3,3,1）。$n_1=7$，$n_2=n_3=5$，$n=17$。算得

$$B_n^{*2} = \frac{1}{72}\{17\times16\times39 - 1398 + (36\times17\times16\times15)^{-1}330\times24$$

$$+ (8\times17\times16)^{-1}82\times28\} = 127.9321,\quad B_n^* = 11.3107$$

$$A_n=\frac{1}{4}(17^2-25-25-49)=47.25.$$

c. 按（4·129）算出 $C=47.25+11.3107\times2.5758=76.384$（此处取 $\alpha=0.01$，$u_{0.01}=2.5758$）此值小于 $R^*(n)=83$，故得出与例4·10一样的结论。若取 $\alpha=0.001$，则本例计算结果在否定域边缘附近，而按例4·10，则与临界值有些距离。显示本方法更为灵敏一些。

二、完全随机区组秩检验

有m个处理要在 n 个区组中进行比较。各区组的大小都是m，可容纳每处理一次且仅一次。区组内各试验单元假定相当均匀，而不同区组则有较大差异（这是区组划分成功的标志）。又假定处理与区组之间并无交互效应。

在传统的正态方差分析中，对这个试验引进统计模型

$$X_{ij}=\mu+\alpha_i+\beta_j+e_{ij},\quad i=1,\cdots,m,\quad j=1,\cdots,n,$$

此处 X_{ij} 为处理 i 在第 j 区组内的试验值。μ,α_i,β_j，e_{ij} 分别为总平均、处理效应、区组效应与随机误差。并假定$\{e_{ij}\}$全体独立，$e_{ij}\sim N(0,\sigma^2)$。

现在我们把这个模型放宽，只假定：

a. 全部 mn 个试验结果独立。

b. 若处理效应不存在，则在每一区组内的m个试验结果同分布。

c. 处理效应大的水平，其试验值倾向于增大（当然，减少也可以，这无关紧要）。此语可确切地解释为：若 $\alpha_u>\alpha_v$，则 $X_{uj}\overset{r}{>}X_{vj}$，对 $j=1,\cdots,n$。

1. Friedman 检验。暂设所有 X_{ij} 的分布函数都处处连续，因而不发生结的问题。

固定j，以 R_{ij} 记 X_{ij} 在$\{X_{1j},\cdots,X_{mj}\}$中的秩。令

$$R_i(n)=\sum_{j=1}^{n}R_{ij}/n,\quad i=1,\cdots,m \tag{4·130}$$

$R_i(n)$是第 i 处理的 n 个试验值，在各自的区组内的秩的平均。据上述假定 c，若处理效应确存在，则对某些 i，$R_i(n)$ 之值将大，而对另一些 i 则小。考虑到所有 $R_i(n)$的平均值为

$$\sum_{i=1}^{m} R_i(n)/m = \frac{1}{2}(m+1).$$

我们可引进下述统计量

$$Q_n = \frac{12n}{m(m+1)} \sum_{i=1}^{m}\left(R_i(n) - \frac{m+1}{2}\right)^2, \tag{4·131}$$

作为衡量处理效应是否存在的指标——Q_n 愈大，处理效应愈像是存在，这导致如下的检验：当

$$Q_n > C \tag{4·132}$$

时，否定“无处理效应”的原假设H。这个检验是 Friedman 在1937年提出来的，通常就冠以他的名字。为确定C，要确定在原假设H之下 Q_n 的分布。对较小的m，n，Friedman 及其他学者给出过这分布，但范围有限。当 n 较大时可使用下面的极限定理：在前述 a、b、c 假定之下，若处理效应不存在，则当 $n\to\infty$时有

$$Q_n \xrightarrow{\mathscr{L}} \chi^2_{m-1}. \tag{4·133}$$

为证明这结果，考虑一串 $m-1$ 维随机向量

$$\xi_j = (R_{1j}, \cdots, R_{m-1,j}),\ j = 1, 2, \cdots$$

根据上述假定 a,b，这一串随机向量独立同分布。其数学期望向量和协方差阵分别为

$$E(\xi_1) = \left(\frac{m+1}{2}, \cdots, \frac{m+1}{2}\right)'.$$

$$\mathrm{Cov}(\xi_1) = (\lambda_{ij}),\ \lambda_{ii} = \frac{m^2-1}{12},\ \lambda_{ij} = -\frac{m+1}{12},\ i \neq j,$$
$$i,\ j = 1, \cdots, m-1.$$

易算出$(\mathrm{Cov}\ (\xi_1))^{-1} = (\rho_{ij})$，其中 $\rho_{ii} = 2\dfrac{12}{m(m+1)}$，$\rho_{ij} = \dfrac{12}{m(m+1)}$当 $i \neq j$．故按中心极限定理，有

$$\eta_n=\left(\sqrt{n}\left(\left(R_1(n)-\frac{m+1}{2}\right),\cdots,\left(R_{m-1}(n)-\frac{m+1}{2}\right)\right)\right)'$$

$$\xrightarrow{\mathscr{L}} N\left(0,\ (\rho_{ij})\right),$$

因而 $\qquad \eta_n'(\rho_{ij})\eta_n \xrightarrow{\mathscr{L}} \chi^2_{m-1}$. $\qquad$ (4·134)

(4·134)左边等于

$$\frac{12n}{m(m+1)}\left\{2\sum_{i=1}^{m-1}\left(R_i(n)-\frac{m+1}{2}\right)^2+\sum_{i\neq j,\,i,j=1}^{m-1}\left(R_i(n)-\frac{m+1}{2}\right)\left(R_j(n)-\frac{m+1}{2}\right)\right\},$$

因为

$$\sum_{i,j=1}^{m-1}\left(R_i(n)-\frac{m+1}{2}\right)\left(R_j(n)-\frac{m+1}{2}\right)$$

$$=\left(\sum_{i=1}^{m-1}-\left(R_i(n)-\frac{m+1}{2}\right)\right)^2=\left(-\left(R_m(n)-\frac{m+1}{2}\right)\right)^2,$$

知

$$\sum_{i\neq j,\,i,j=1}^{m-1}\left(R_i(n)-\frac{m+1}{2}\right)\left(R_j(n)-\frac{m+1}{2}\right)$$

$$=-\sum_{i=1}^{m-1}\left(R_i(n)-\frac{m+1}{2}\right)^2+\left(R_m(n)-\frac{m+1}{2}\right)^2,$$

于是得到(4·134)左边正好就是 Q_n。这证明了所要的结果。据此结果，当 n 较大时，(4·132)中的常数 C 近似地可取为

$$C=\chi^2_{m-1}(\alpha), \qquad (4\cdot135)$$

此处 α 为给定的检验水平。

在正态方差分析中，常用于检验"处理效应为0"的检验是以

$$n\ (n-1)\sum_{i=1}^{m}(\overline{X}_i-\overline{X})^2\Big/\sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij}-\overline{X}_i)^2>F_{m-1,\,m(n-1)}(\alpha)$$

为否定域的 F 检验 $\mathscr{F}$. Friedman 检验(FR)与此检验的对比如何？可以证明：在如下型的对立假设

$$X_{ij}\sim F(x-\alpha_i-\beta_j),\ i=1,\cdots,m,\ \ j=1,\cdots,n \qquad (4\cdot136)$$

之下，二者的渐近相对效率为

$$\text{ARE}(FR,\ \mathscr{F};\ F)=\frac{m}{m+1}\text{ARE}(W,t;\ F),$$

其中W,t 分别是 Wilcoxon 两样本秩和检验与两样本 t 检验.由此式可知，当m不太小时，Friedman 检验与F检验相比处在有利地位，当m小时则否。

2．结存在的情况

结存在的情况比较重要，因为，有时区组试验的观察结果是属性的，例如产品的等级。在这种情况下通常的 F 检验不能用，而 Friedman 检验只须稍加修改就可以（这一点也适用于多样本问题，见前）。修改步骤如下：

a．把 R_{ij} 在各区组内按平均法修改为 R^*_{ij}，而 $R_i(n)$修改为 $R^*_i(n)=\sum_{j=1}^{n}R^*_{ij}/n$.在 Q_n 中以 $R^*_i(n)$代 $R_i(n)$，所得结果记为 Q^*_n。

b．对每个j，$j=1,\cdots,n$，找出$(X_{1j},\cdots,X_{mj})$的结统计量$(\tau_{1j},\cdots,\tau_{g_j j})$。计算

$$\widetilde{Q}_n=\left\{1-\sum_{j=1}^{n}\sum_{r=1}^{g_j}(\tau^3_{rj}-\tau_{rj})\Big/\Big(nm(m^2-1)\Big)\right\}^{-1}Q^*_n,\tag{4·137}$$

可以证明：在原假设成立之下，当 $n\to\infty$时有 $\widetilde{Q}_n\xrightarrow{\mathscr{L}}\chi^2_{m-1}$。

故以$\{\widetilde{Q}_n>\chi^2_{m-1}(\alpha)\}$为否定域的检验，有渐近水平 α。

3．对立假设有序的情况

设根据问题的实际背景，有理由认为：当原假设（无处理效应)不成立时，水平编号愈大者愈优.这可以解释为：若以F_{ij} 记 X_{ij} 的分布，则对于任何固定的j，当 $u<v$ 时 $F_{vj}(x)\leqslant F_{uj}(x)$，即

$$X_{vj}\overset{\tau}{>}X_{uj},\ 1\leqslant u<v\leqslant m,\ j=1,\cdots,n.\tag{4·138}$$

仍以 $R_i(n)$记 $X_{i1},\cdots,X_{in}$ 在各自的区组里的秩的和。若

（4·138）成立，则 i 愈大时，$R_i(n)$ 且倾向于大。由于$R_1(n)+\cdots+R_m(n)=\frac{1}{2}nm(m+1)$ 是一常数，知统计量

$$T_n=\sum_{i=1}^{m} iR_i(n) \tag{4·139}$$

也倾向于取较大之值，这导出如下的检验：

当 $T_n>C$ 时否定原假设。　　　　（4·140）

不难证明（留给读者作为练习）：设 X_{ij} 的分布连续，则在原假设成立之下，当 $n\to\infty$时有

$$(T_n-a_n)/b_n \xrightarrow{\mathscr{L}} N(0,1), \tag{4·141}$$

其中$a_n=\frac{1}{4}nm(m+1)^2$，$b_n^2=\frac{1}{12}\sqrt{n(m-1)}m(m+1)$。利用（4·141），当 n 较大时，（4·140）中的C可近似地取为 $a_n+b_nu_\alpha$，α 为给定的检验水平。

4．另一种检验方法

前已指出，当m较小时，Friedman 检验的表现不理想。其原因何在，对$m=2$ 的情况稍加分析不难看出其端倪。简单计算表明当$m=2$ 时，Friedman 统计量（4·131）有 $2n(A/n-1/2)^2$的形式，其中A为处理 1 在各区组中取秩 1 的次数。因此，在这个场合下 Friedman 检验事实上就是符号检验，而一般说来，符号检验的表现不如更精细些的秩检验，例如 Wilcoxon 检验。这从（4·79）式下面那个表中的对比可看出一些。

进一步看，Friedman 检验上述缺点的根子在于，它在定秩时只利用了各区组内的相互比较。诚然，在区组有较大差异时，简单地把全部X_{ij}混在一起定秩（如在多样本问题中的做法）不行。但如先对样本 X_{ij} 作一些处置以除去区组差异的影响，然后合在一起定秩，则道理上说得过去。1962年，Hodges 和 Lehmann 根据这个想法，提出了一个效率更高的秩检验。

以 $X_{\cdot j}$ 记区组 j 的平均值：$X_{\cdot j}=\sum_{i=1}^{m}X_{ij}/m$。条件比较好的区组 j，$X_{\cdot j}$ 倾向于大些。把区组 j 内的观察值 $X_{1j},\cdots,X_{mj}$ 都减去这区组平均值，得

$$X'_{ij}=X_{ij}-X_{\cdot j},\ i=1,\cdots,m,\quad j=1,\cdots,n$$

在样本 $\{X'_{ij}\}$ 中，由区组差异带来的影响已消去了。

现以 R'_{ij} 记 X'_{ij} 在合样本 $\{X'_{uv}:u=1,\cdots,m,\ v=1,\cdots,n\}$ 中的秩，然后按多样本问题的方式去处理。即引进定义在 $\{1,2,\cdots,mn\}$ 上的计分函数 a。计算

$$S_i=\sum_{j=1}^{n}a(R'_{ij}),\ i=1,\cdots,m$$

然后按（4·109）式计算统计量(注意此处相当于 $n_1=n_2=\cdots=n_m=n$ 的情况)

$$T_n=n(mn-1)\sum_{i=1}^{m}(S_i/n-\bar{a})^2/\sum_{i=1}^{mn}\left(a(i)-\bar{a}\right)^2,\quad (4\cdot142)$$

其中 $\bar{a}=\sum_{i=1}^{mn}a(i)/mn$。可以证明，在 X_{ij} 的分布连续的条件下，当原假设成立时有

$$T_n\xrightarrow{\mathscr{L}}\chi^2_{m-1},\qquad n\to\infty,\qquad (4\cdot143)$$

因此，以

$$\{T_n>\chi^2_{m-1}(\alpha)\}\qquad (4\cdot144)$$

为否定域的检验，当 n 较大时其水平接近 α。

特别，在 $a(i)=i$ 时，（4·142）有形式

$$T_n=\frac{12}{m(mn+1)}\sum_{i=1}^{m}\left(R'_{i\cdot}-\frac{1}{2}(mn+1)\right)^2,\ R'_{i\cdot}=\sum_{j=1}^{n}R'_{ij}/n,\qquad (4\cdot145)$$

此检验对 Friedman 检验的 ARE，相当于 Wilcoxon 检验对符号检验的 ARE。

如果在经过变换以后的样本 $\{X_{ij}':i=1,\cdots,m,\ j=1,\cdots,n\}$ 中有结存在，则可以按照多样本问题中修正统计量（4·109）的方式，去修正（4·142）。修正后的统计量在原假设成立时，仍

有(4·143).

§4·4 随机性与独立性的秩检验

一、随机性的秩检验

1. 问题提法 设有一维样本 $X_1,\cdots,X_n$. 所谓"随机性假设".是指

$H: X_1,\cdots,X_n$ 是从某总体中抽出的简单样本 (4·146)

这个原假设。在一些情况下，事先已知或有理由假定 $X_1,\cdots,X_n$ 独立。这时随机性假设归结为"$X_1,\cdots,X_n$ 同分布".这是多样本问题当 $n_1=\cdots=n_n=1$ 时的一个特例.

容易理解：随机性假设必须针对特定的对立假设去检验，才是有意义的问题。可以举几个例子来说明这一点.

a. 设 $X_1=-1$, $X_2=0$, $X_3=-2$, $X_4=10000$

初一看，这很不符合随机性假设。因为其中一个样本跑到了离群很远的地方。但如设想总体分布是：$P(X=-2)=P(X=-1)=P(X=0)=P(X=10000)=1/4$，则把这组样本认为是从这分布中抽取的，就显得很自然.反之，若事先已知各样本都服从方差为1的正态分布(这时随机性假设等于各样本的期望相同)，则上述样本与随机性，看来相去甚远.

b. 设有样本 $X_1,\cdots,X_n$，满足条件 $X_1<X_2<\cdots<X_n$.初一看样本因有一种上升的趋势，而与随机性不合. 可是，在已有了 $X_1,\cdots,X_n$ 这 n 个数值后，当随机性成立时，它的任意一种排列方式(一共有 $n!$种排列方式)有同等的可能 $1/n!$.换句话说，表面上看来上升的 $X_1,\cdots,X_n$，其实与其他任何排列比，毫无特异之处。但如我们在事先有理由认为 X_i 满足线性回归模型

$X_i=\alpha+\beta t_i+e_i,\quad i=1,\cdots,n, e_1,\cdots, e_n\ iid.$ (4·147)

且 $t_1<t_2<\cdots<t_n$.则 $\beta=0$ 相应于随机性。如我们针对对立假设 $\beta>0$，则 $X_1<\cdots<X_n$ 这样的样本，就显得与 $\beta>0$ 很合拍，

因而就有理由怀疑随机性假设不成立。

c. 设样本只能取0,1两个值，而我们得到样本0,0,1,1,1。初一看也觉得这与随机性不甚合拍。因为0,1各自聚在一起，像是有某种相关性存在。可是与在b中一样，在随机性成立的前提下，两个0三个1的10种可能的排列形状有完全同样的概率$\frac{1}{10}$因此，0,0,1,1,1这个结果与(比方说)1,0,1,1,0相比，无任何特异之处，而1,0,1,1,0这个结果看上去像是符合随机性。但如我们有理由认为，若不随机则是由"正相关"所引起的，则0,0,1,1,1这个结果就像是与对立假设更接近些.

我们下面要讨论的随机性秩检验，就是针对这里的b、c两种情况。

2. 针对上升趋势的秩检验.我们假定$X_1,\cdots,X_n$独立X_i的分布F_i连续，$i=1,\cdots,n$。随机性假设相当于$F_1=F_2=\cdots=F_n$，而对立假设为

$$X_n \overset{r}{>} X_{n-1} \overset{r}{>} \cdots \overset{r}{>} X_1, \tag{4·148}$$

以R_i记X_i在$X_1,\cdots,X_n$中之秩.若对立假设(4·148)正确，则X_i倾向于排在第i位附近，即R_i倾向于取i附近的值。因此，统计量

$$\begin{aligned}T_n &= \sum_{i=1}^{n}(R_i-i)^2=\sum_{i=1}^{n}R_i^2+\sum_{i=1}^{n}i^2-2\sum_{i=1}^{n}iR_i \\ &= \frac{1}{3}n(n+1)(2n+1)-2\sum_{i=1}^{n}iR_i \end{aligned}\tag{4·149}$$

倾向于小。记$S_n=\sum_{i=1}^{n}iR_i$，得到如下的检验：

$$\text{当 } S_n>C \text{ 时否定原假设。} \tag{4·150}$$

对较小的n,S_n在原假设下的分布易于定出,利用它,可根据给定的检验水平α确定(4·150)中的临界值C.当n较大时，可使用下面的极限定理：当$X_1,\cdots,X_n$独立同分布且分布连续时，$n\to\infty$时有

$$\left(S_n-\frac{1}{4}n(n+1)^2\right)\Big/\left(n(n+1)\sqrt{n-1}/12\right)\xrightarrow{\mathscr{L}}N(0,1)$$

$$(4\cdot151)$$

而定出（4·150）中C的近似值

$$C=\frac{1}{4}n(n+1)^2+n(n+1)\sqrt{n-1}\,u_\alpha. \qquad (4\cdot152)$$

（4·151）可从定理4·4推出：取$C_{ni}=i$，$i=1,\cdots,n$而$\varphi(u)=u$，$0<u<1$。

此检验可用于检验模型(4·147)中的$\beta=0$，对立假设是$\beta>0$，在此假定了$t_1<t_2<\cdots<t_n$。在通常的线性回归分析中假定误差e_i服从正态分布$N(0,\sigma^2)$，$\beta=0$的假设是用由最小二乘法导出的t检验去检验之。Staurt在1954—1956年的工作中，考虑了秩检验（4·150）对这个t检验的ARE。在$t_i=i$的情况，结果为$\sqrt{3/\pi}\approx0.98$，很接近于1。一般地，当误差e_i独立同分布，且公共分布F有密度时，有ARE（(4·150)，t，F）＝（ARE（W t，F））$^{1/2}$。这里，后一表达式中的W，t分别是两样本问题的Wilcoxon检验和t检验。

由此看出：即使一个表面上看来相当粗糙的检验（4·150），也对传统的t检验有很高的竞争力。甚至在正态场合(这时t检验处在优越地位）也是如此。这使我们对非参数方法的效力具有信心。也可以反过来去看：通过引进很细致的模型和分析方法（正态假定，最小二乘法，t分布等)，比之直接从常识出发而导出的方法，并未增加多少东西。

3．针对相关性的对立假设——游程检验

在一个只包含两个符号的序列中，由相邻同一符号形成的一段叫一个游程。例如在1001110100011中，有4个“1游程”，即1，111，1，11；3个“0游程”，即00，0，000。一共有7个游程。

先设样本$X_1,\cdots,X_n$都只取0、1两个值。则每组试验结果，都是一个由0，1构成的序列。如果$X_1,\cdots,X_n$有随机性（独立同

分布)，则在这序列中 0，1 两个符号应当既不太集中又不太分散，因此游程总数应当适中(不大不小)。反之，若相邻变量之间存在正相关，则 $X_{i-1}=1$ 易引起 $X_i=1$。这时 0，1 在序列中会倾向于更集中，而导致游程总数减少。类似地，若相邻变量之间存在负相关，则游程总数倾向于增多。这样，序列中的游程总数 ξ 提供了随机性的一种检验方法。

不难证明(见 Feller：《An Introduction to Probability Theory and its Applications》第二章)：若有 m_1 个 0 和 m_2 个 1 随机地排成一列(意思是：这 m_1+m_2 个符号的 $(m_1+m_2)!$ 种可能的排列方式为同等可能)，并以 ξ 记 1 游程个数，则

$$P(\xi=k)=\binom{m_1+1}{k}\binom{m_2-1}{k-1}\Big/\binom{m_1+m_2}{m_1} \qquad (4\cdot153)$$

因为 0、1 游程个数相差至多为 1，可以就使用 ξ 来构造检验，而不必一定使用游程总数。

有了样本 $X_1,\cdots,X_n$ 后，先数出其中 0 的个数 m_1，1 的个数 $m_2=n-m_1$。就这个 m_1，m_2，算出对各个 k 的概率(4·153)之值。如针对的对立假设是正相关，则要取 ξ 的小值作为否定域：

当 $\xi<C$ 时否定 (4·154)

C 根据条件

$$\sum_{k=1}^{c-1}\binom{m_1+1}{k}\binom{m_2-1}{k-1}\Big/\binom{n}{m_1}=\alpha \qquad (4\cdot155)$$

去选择。当这样的 C 不存在时，可适当调整 α 之值，或施行随机化。

如针对的对立假设为负相关，则应取 $\{\xi>C\}$ 为否定域。C 的决定法与(4·155)相似，只和号改为 $\sum\limits_{k=c_1+1}^{m'}$ $m'=\min(m_2, m_1+1)$。若对立假设兼有正负相关之可能(双侧性)，则应取 $\{\xi<C_1\}\cup\{\xi>C_2\}$ 为否定域，C_1、C_2 由

$$\sum_{k=1}^{c_1-1}\binom{m_1+1}{k}\binom{m_2-1}{k-1}\Big/\binom{n}{m_1}=\frac{\alpha}{2}=\sum_{k=c_2+1}^{m'}\binom{m_1+1}{k}$$

$$\cdot\binom{m_2-1}{k-1}\Big/\binom{n}{m_1} \qquad (4\cdot156)$$

确定。

当 m_1 和 m_2 都较大时，上述利用精确分布的做法不可行。这时可使用 Mann 和 Wald 在 1940 年证明的一个关于 ξ 的极限定理：若 $m_1\to\infty$，$m_2\to\infty$ 而 m_1/m_2 和 m_2/m_1 始终保持有界，则有

$$\frac{(m_1+m_2)^{3/2}}{m_1m_2}\left(\xi_1-\frac{m_1m_2}{m_1+m_2}\right)\xrightarrow{\mathscr{L}}N(0,1). \qquad (4\cdot157)$$

根据这个定理，当 m_1，m_2 都较大时，(4·154) 中的 C 可近似地用

$$C=\frac{m_1m_2}{m_1+m_2}-\frac{(m_1+m_2)^{3/2}}{m_1m_2}u_\alpha \qquad (4\cdot158)$$

代替之。对 (4·156) 中的 C_1,C_2，也有类似近似式。

现在考虑样本取任意值的一般情况。它仍是基于类似的想法当有正相关时，序列 $X_1,\cdots,X_n$ 中，小值倾向于扎堆，大值也如此。故若以 $\widehat{X}$ 记 $X_1,\cdots,X_n$ 的样本中位数，而令 $X_i'=0$ 或 1，视 $X_i\leqslant\widehat{X}$ 或 $X_i>\widehat{X}$ 而定，则全由 0，1 组成的序列 $X_1',\cdots,X_n'$ 中，0（相应于 $X_1,\cdots,X_n$ 中的小值）倾向于成堆，1 也一样。换言之，1 游程的个数 ξ 倾向于减少。当相邻变量为负相关时，情况类似。因此，在得到序列 $X_1',\cdots,X_n'$ 后，即用前面处理 0，1 序列的方法去检验之即可。

二、独立性的秩检验

设 $(X_1,Y_1),\cdots,(X_n,Y_n)$ 是二维随机向量 (X,Y) 的简单样本，要检验假设

$$H: X,Y\text{独立}, \qquad (4\cdot159)$$

这个问题在例 3·9 中讨论过。在那里我们引进了 Kendall 的 τ 检验。不难看出，该检验的检验统计量事实上只与 X_i 和 Y_i 的秩有关，因而是一个秩检验。本段再介绍几个与此问题有关的秩检验。

1. Spearman 的秩相关检验.

Spearman 在 1904 年引进的秩相关检验，属于历史上秩方法最早的应用之一。这方法的概念很简单。先考虑结不存在的情况。以 Q_i 记 X_i 在 $X_1,\cdots,X_n$ 中之秩，R_i 记 Y_i 在 $Y_1,\cdots,Y_n$ 中之秩。用"秩样本"$(Q_1,R_1),\cdots,(Q_n,R_n)$ 代替原样本计算相关系数

$$r_n=\sum_{i=1}^{n}(Q_i-\bar{Q})(R_i-\bar{R})/\left(\sum_{i=1}^{n}(Q_i-\bar{Q})^2\sum_{i=1}^{n}(R_i-\bar{R})^2\right)^{1/2} \tag{4·160}$$

其中 $\bar{Q}=\dfrac{n+1}{2}=\bar{R}$。然后，视对立假设为"正相关"、"负相关"、或"正负相关都可能"，而相应地取否定域

$$\{r_n>C'\},\{r_n<C'\},\text{或}\{|r_n|>C'\}. \tag{4·161}$$

用秩相关系数(4·160)代替通常的相关系数来作独立性检验，理由在于在原假设下当分布连续时，秩相关系数有"分布无关"性。事实上，由于 $\bar{Q}=\bar{R}=\dfrac{n+1}{2}$，而$\{Q_1,\cdots,Q_n\}$及$\{R_1,\cdots,R_n\}$都取 $1,2,\cdots,n$ 一次且仅一次，有$\sum\limits_{i=1}^{n}(Q_i-\bar{Q})^2=\sum\limits_{i=1}^{n}(R_i-\bar{R})^2=\sum\limits_{i=1}^{n}\left(i-\dfrac{n+1}{2}\right)^2=\dfrac{n}{12}(n^2-1)$，是一个只依赖 n 的常数。又

$$\sum_{i=1}^{n}(Q_i-\bar{Q})(R_i-\bar{R})=\sum_{i=1}^{n}Q_iR_i-n\bar{Q}\bar{R}=L_n-\frac{n}{4}(n+1)^2,$$

其中$L_n=\sum\limits_{i=1}^{n}Q_iR_i$。这样，要找 r_n 的分布，只须找 L_n 的分布。但易见当原假设(X，Y独立)成立时，有

$$L_n\overset{d}{=}\sum_{i=1}^{n}iR_i, \tag{4·162}$$

这可如下证明：固定$(Q_1,\cdots,Q_n)$的一组值$(q_1,\cdots,q_n)$，而去考虑在$(Q_1,\cdots,Q_n)=(q_1,\cdots,q_n)$的条件下，$L_n$的条件分布。

$q_1,\cdots,q_n$ 必须取 $1,\cdots,n$ 中各数一次且仅一次。由于 X，Y 独立，知 $(Q_1,\cdots,Q_n)$ 与 $(R_1,\cdots,R_n)$ 独立（因为前者只与 $X_1,\cdots,X_n$ 有关，而后者只与 $Y_1,\cdots,Y_n$ 有关）。故上述条件分布，也就等于 $\sum_{i=1}^{n} q_i R_i$ 的无条件分布。由于 $q_1,\cdots,q_n$ 跑遍 $1,\cdots,n$，可找到 $(1,\cdots,n)$ 的一个置换 $(l_1,\cdots,l_n)$，使 $q_{l_i}=i$，$i=1,\cdots,n$，这样 $\sum_{i=1}^{n} q_i R_i=\sum_{i=1}^{n} iR_{l_i}$。但因 $R_1,\cdots,R_n$ 为简单样本 $Y_1,\cdots,Y_n$ 的秩，有 $(R_{l_1},\cdots,R_{l_n})\overset{d}{=}(R_1,\cdots,R_n)$。故知 $\sum_{i=1}^{n} q_i R_i=\sum_{i=1}^{n} iR_{l_i}$，从而证明了 L_n 的分布与 $\sum_{i=1}^{n} iR_i$ 的分布相同，而后者在原假设下为分布无关。

根据上述，否定域（4·161）可以相应地用下述否定域取代：

$$\{L_n>C\},\{L_n<C\},\text{或}\{|L_n-\frac{n}{4}(n+1)^2|>C\} \quad (4\cdot163)$$

为确定临界值 C，当 n 较小时可直接计算 $\sum_{i=1}^{n} iR_i$ 的精确分布。在 n 较大时可利用 L_n 的渐近正态性。据上述，这就是（4·151）（把其中的 S_n 改为 L_n）。

Spearman 秩相关检验可以从另一个考虑得到。如果 X，Y 为正相关，则 $(Q_1,\cdots,Q_n)$ 与 $(R_1,\cdots,R_n)$ 应为"同步"，即当 Q_i 小（大）时 R_i 也倾向于小（大）。因此，表达式 $\sum_{i=1}^{n}(Q_i-R_i)^2$ 应倾向于小。若 X，Y 为负相关，则 Q_i 小（大）时 R_i 倾向于大（小），这时 $\sum_{i=1}^{n}(Q_i-R_i)^2$ 倾向于大，然而

$$\sum_{i=1}^{n}(Q_i-R_i)^2=\sum_{i=1}^{n} Q_i^2+\sum_{i=1}^{n} R_i^2-2\sum_{i=1}^{n} Q_i R_i$$
$$=\frac{1}{3}n(n+1)(2n+1)-2\sum_{i=1}^{n} Q_i R_i,$$

故基于 $\sum_{i=1}^{n}(Q_i-R_i)^2$ 的检验法，与基于 $L_n=\sum_{i=1}^{n}Q_iR_i$ 的检验法完全相同。

此检验的效率如何？拿它与常见的相关系数检验法，在正态分布 $N(a,b,\sigma_1^2,\sigma_2^2,\rho)$的场合下去比较，$\rho=0$ 相应于X,Y独立，$\rho>0$ 和 $\rho<0$ 分别为正、负相关。可以证明：针对这个场合，Spearman 检验对通常相关系数的检验的 ARE 为 $9/\pi^2\approx0.912$。这是 Kruemer 在 1974 年证明的。这里我们又得到鲜明的印象：尽管 Spearman 检验看上去很粗糙，但它与建立在很特殊的假定（正态）之下，并经过很复杂的分析才得出的检验法对比，效率损失其实很小。而 Spearman 检验还有“分布无关”的优点，即不致因模型假定错误而发生大问题。

如果结存在，则需要作适当的修正。我们从统计量 $\sum_{i=1}^{n}(Q_i-R_i)^2=T_n$ 出发（上面已指出这与 Spearman 的秩相关系数等价）去修正更方便些。步骤如下：

a．在$(X_1,\cdots,X_n)$和$(Y_1,\cdots Y_n)$各自的范围内，用平均法定秩，把 Q_i,R_i 分别修正为 Q_i^*,R_i^*，把 T_n 修改为 $\sum_{i=1}^{n}(Q_i^*-R_i^*)^2=T_n^*$

b．定出$(X_1,\cdots,X_n)$和$(Y_1,\cdots,Y_n)$各自的结统计量$(\tau_1,\cdots,\tau_p)$和$(\eta_1,\cdots,\eta_q)$。计算

$$A_n=\frac{1}{6}(n^3-n)-\frac{1}{12}\left(\sum_{i=1}^{p}(\tau_i^3-\tau_i)+\sum_{i=1}^{q}(\eta_i^3-\eta_i)\right),$$

$$B_n^2=\frac{1}{36}(n-1)n^2(n+1)^2\left(1-\frac{\sum_{i=1}^{p}(\tau_i^3-\tau_i)}{n^3-n}\right)\cdot\left(1-\frac{\sum_{i=1}^{q}(\eta_i^3-\eta_i)}{n^3-n}\right).$$

c．经过上述修正后，在原假设成立之下有

$$(T_n^* - A_n)/B_n \xrightarrow{\mathscr{L}} N(0,1), \qquad (4 \cdot 164)$$

据（4·164），当 n 较大时，针对三个对立假设："正相关"、"负相关"和"正、负相关都可能"，否定域可依次取为

$$T_n^* < A_n - B_n u_\alpha, \quad T_n^* > A_n + B_n u_\alpha, \quad 及 |T_n^* - A_n| > B_n u_{\alpha/2}$$

$$(4 \cdot 165)$$

此处 α 为给定的检验水平。

可以引进一个在集合 $\{1,2,\cdots,n\}$ 上非降的计分函数 $a(\cdot)$，而把 Spearman 秩相关检验推广为：令 $\tilde{L}_n = \sum_{i=1}^{n} a(R_i)a(Q_i)$。在结不存在及原假设成立，且 $a(\cdot)$ 满足一定的条件（例如，$a(\cdot)$ 是定理 4·4 或定理 4·5 中的那种形式）时，有

$$\sqrt{n-1}\,(\tilde{L}_n - n\bar{a}^2) \Big/ \sum_{i=1}^{n} (a(i) - \bar{a})_2 \xrightarrow{\mathscr{L}} N(0,1), \qquad (4 \cdot 166)$$

此处 $\bar{a} = \sum_{i=1}^{n} a(i)/n$。这样，就可基于 $\tilde{L}_n$ 作出针对前述三种对立假设的大样本秩检验。（4·166）的证明方法与 $a(i) = i$ 时一样，细节留给读者。

不同 $a(\cdot)$ 的取法，导致在种种特殊的对立假设上表现不一的检验。例如，若取 $a(i) = \Phi^{-1}\left(\frac{i}{n+1}\right)$ 或 $a(i) = E\xi_{ni}$，即两样本问题中 Van der Waerden 检验和 Fisher Yates 检验中那种取法，则在正态对立假设 $N(a, b, \sigma_1^2, \sigma_2^2, \rho)$ 上，这检验对通常的相关系数检验的 ARE 为 1。

2. 列联表的秩处理。如果 X，Y 都只取有限个数值（也可能，X，Y 都是属性变量，可加以数量化），则 X，Y 独立性的检验，通常在列联表中用 χ^2 检验法进行，用我们前面讲过的多样本检验和秩相关检验，可以把秩方法引进来处理这种列联表。我们先把数据表为列联表的形式（这里 $a_1, \cdots, a_r$ 两两不同，$b_1, \cdots, b_s$ 两两不同。它们也可以只是属性变量中的等级符号）。

Y \ X	b_1	b_2	…	b_j	…	b_s	行和
a_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1s}	n_1
a_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2s}	n_2
…	…	…	…	…	…	…	…
a_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{is}	n_i
…	…	…	…	…	…	…	…
a_r	n_{r1}	n_{r2}	…	n_{rj}	…	n_{rs}	n_r
列和	d_1	d_2	…	d_j	…	d_s	n

先考虑对立假设无方向的情况。或简单地说，检验是针对一切对立假设。这时，可以把上表中每一行的数据看作为从一个总体中抽出的样本，一共有 r 组样本。在 X，Y 独立时，给定 X 时 Y 的条件分布即等于 Y 的无条件分布，因而上述 r 组样本所来自的 r 个总体，有同一的分布。这可以用结存在时的多样本检验法去检验之。例如，用 Kruskal-Wallis 检验，其有结存在的情况为公式(4·120)。

从表上看出合样本的结统计量为$(d_1,\cdots,d_s)$。现在算第 i 组(即第 i 行)样本的修正后的秩平均。第1列内每个样本的(平均)秩为$\frac{1}{2}(d_1+1)$，第2列为 $d_1+\frac{1}{2}(d_2+1)$，…，一般，第 j 列为 $d_1+\cdots+d_{j-1}+\frac{1}{2}(d_j+1)$。由此可知

$$R_i^*=\sum_{j=1}^{s}n_{ij}\left(\sum_{u=1}^{j}d_u-\frac{d_j-1}{2}\right)\Big/n_i,\quad i=1,\cdots,r,$$

以这个 R_i，以及 $q=s$，$\tau_j=d_j$，$j=1,\cdots,s$ 代入(4·120)，得检验统计量

$$T^*=\left(1-\frac{\sum_{j=1}^{s}(d_j^3-d_j)}{n^3-n}\right)^{-1}\frac{12}{n(n+1)}\sum_{i=1}^{r}n_i\left(R_i^*-\frac{n+1}{2}\right)^2,\tag{4·167}$$

水平 α 大样本否定域为$\{T_n^*>\chi^2_{r-1}(\alpha)\}$。

在列联表上，Y之值 $b_1,\cdots,b_s$ 我们假定已按由小到大的次序排列。若Y为属性变量，则随意排一个次序都可以。不论你如何排，所得出的统计量(4·167)在原假设下总有极限分布 χ^2_{r-1},因而检验的方式不变，但是，Y的值排序不一样时，由之算出的R_i^* 值也不一样，故统计量(4·167)在一组特定的样本之下所取的值，与Y值的排序有关。这一来就有如下的可能：甲、乙两人面对同一组试验结果，因甲、乙对Y值排序不同，甲所算出的T_n^* 值超过 $\chi^2_{r-1}(\alpha)$，而乙算出的则否。这样甲、乙两人就得出不同的结论，甲否定原假设而乙接受。为免除这个不便，可在事先商定一种次序，一般可按就某种指标而言是自然的次序。例如，Y这个变量是关于一个患者得某病的程度。它可以自然地按由重到轻排序为重度、中度、轻度及无病四种，或反过来也可以。

如果要构造一个针对有序的对立假设，则必须先对Y之取值排定一种有意义的次序，如上文中按病情严重程度排序。这个做了以后，就要确定X的取值的一个次序，使得当对立假设为真时，X的大值所相应的Y总体(其样本是列联表的一行)，随机地大于X的小值所相应的Y总体。然后，按多样本问题对立假设有序且结存在时的公式(见例 4·11 前面的讨论)去处理即可。

针对有序对立假设的另一种处理方法是利用 Spearman 秩相关检验中的想法。每一个样本，按X值的次序可定出一个秩Q_{ij}，按Y值次序可定出一个秩 R_{ij}。然后按 Spearman 检验有结存在时的方法去处理即可。具体如下。设在前面的列联表中，X值的次序已按 $a_1<\cdots<a_r$ 排列，而Y值按 $b_1<\cdots<b_s$ 排列。在第 i 行 j 列中之任一样本。按X值排序，其平均秩应为 $n_1+\cdots+n_{i-1}+\frac{1}{2}(n_i+1)=(n_1+\cdots+n_i)-\frac{1}{2}(n_i-1)$，此即 Q_{ij}（实际与 j 无关），而按Y值排序，其平均秩应为（$d_1+\cdots+d_j$）－

$\frac{1}{2}(d_j+1)$，此即 R_{ij}（实际与 i 无关）。于是，在 Spearman 检验有结情况的 a、b、c三个步骤中，步骤 a 中之统计量 T_n^* 为

$$T_n^* = \sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}\left\{\left(d_1+\cdots+d_j-\frac{1}{2}(d_j+1)\right)-\left(n_1+\cdots+n_i-\frac{1}{2}(n_i+1)\right)\right\}^2, \tag{4·168}$$

步骤 b 为计算

$$A_n=\frac{1}{6}(n^3-n)-\frac{1}{12}\left(\sum_{i=1}^{r}(n_i^3-n_i)+\sum_{j=1}^{s}(d_j^3-d_j)\right), \tag{4·169}$$

$$B_n^2=\frac{1}{36}(n-1)n^2(n+1)^2\left(1-\frac{\sum_{i=1}^{r}(n_i^3-n_i)}{n^3-n}\right)\cdot\left(1-\frac{\sum_{j=1}^{s}(d_j^3-d_j)}{n^3-n}\right),$$

然后据（4·164）而得到大样本否定域（4·165）。

看一个数字例子。

例4·12 为检验A、B、C三种药物对治疗某种疾病的效果有否差异，对 167 位患者作了试验．效果(病人治疗后的状况)分严重，中度，轻度，痊瘉四级，以下分别用 1 、 2 、 3 、 4 记之，全部试验结果列表如下：

效果（Y） 药物（X）	1	2	3	4	行和
A	8	8	19	35	70
B	2	3	5	20	30
C	3	4	15	45	67
列和	13	15	39	100	167

先按无序对立假设去检验，用公式(4·167)，算出 $T_n^*=5.501$，

查表知。$P(\chi_2^2>5.501)=0.064$。此值比 0.05 大，故按 $\alpha=0.05$ 的检验水平，尚不能否定"药物与疗效无关"的原假设。

现针对有序对立假设去检验、设根据事先拥有的知识可认为：若各药物疗效不同，则应是

$$\{A\text{比}B\text{差},\ B\text{比}C\text{差(二者至少成立其一)}\}$$

这样的情况，按表中数据，算出（4·168）定义的 T_n^* 取值为 521459.5，而由（4·169）和（4·170）定义的 A_n 和 B_n 分别为 631606 和 48955.79。于是

$$(T_n^*-A_n)/B_n=-2.250,$$

而 $\Phi(-2.250)=0.0123$（Φ 为 $N(0,1)$ 的分布函数），故即使取水平 $\alpha=0.01$，用本法据所得数据，也接近于否定原假设. 很明显，这是因为前法要面向四方的对立假设，因而比较保守，不易发现差异之故。本例也可按通常的列联表 χ^2 检验去做，结果更为保守：算出的 χ^2 值为 6.358，自由度 $(3-1)(4-1)=6$，而 $P(\chi_6^2>6.358)=0.384$。

§4·5 秩方法用于估计问题

非参数方法，尤其是在其较早期的发展中，重点在假设检验。至于估计问题，你当然也可以说，像用样本均值估计一个未知总体的数学期望这类方法，属于非参数估计法。但这类方法，也多半只是把参数统计中习知的问题中的习知的方法，平行地移过来而已，没有多少典型的"非参数"成份。

这原因不难理解。在检验问题中，如果模型中的分布族很大，则为得到在原假设下为分布无关的检验统计量，就必须使用特殊的统计量，它们只依赖于样本中"一般"信息（如次序、秩之类）。例如为检验对称分布 $F(x-\theta)$ 的对称中心 $\theta=0$。若已知 F 为正态，可用通常的一样本 t 检验，但如 F 可为任何对称连续分布，则 t 统计量在 $\theta=0$ 时已非分布无关，而只有像符号统计

量，Wilcoxon 符号秩和统计量等才有此性质，因此，发展合用的非参数检验法是一种必需．

估计问题则不然，特别是点估计，对它并无上述"分布无关"的要求．如在上述例中，当F为正态时，一般用样本均值$\overline{X}$去估计θ．当F可为任意对称分布时，只要F的期望有限，用$\overline{X}$估计θ这个方法仍可用，至于$\overline{X}$在$\theta=0$时并非分布无关这一点，并不影响其应用．又如在两个总体X、Y的分布分别为$F(x)$及$F(x-\theta)$的情况，你可以用正态情况下的估计量$\overline{Y}-\overline{X}$去估计$\theta$，在一般应用问题中，分布$F$往往有良好的性质，如方差有限等，这估计$\overline{Y}-\overline{X}$因之也有较优良的性能，但如要检验$\theta=0$则完全是另一个问题，你不能把通常的$t$检验照搬过来．

虽然如此，估计问题在非参数统计中仍占有重要的地位，且就近年发展看，估计问题在研究工作中还受到更大的重视．这主要是由于，在一些非参数模型中，被估计的量在传统的参数统计中并无可直接类比之物，对这种量的估计，就无法直接搬用参数统计中惯用的方法．第六章中讨论的概率密度估计与非参数回归估计是一个重要的例子．除此以外，非参数检验统计量的发展也提供了一些估计问题的新的处理方法，其优越性与传统方法相比有竞争力．本节中要讨论的对称中心及位置参数的估计问题，就属于这种情况．

一、对称中心的区间估计．

设$X_1,\cdots,X_n$为抽自分布$F(x-\theta)$的简单样本，$F(x)$为关于原点对称的连续分布，其他无所知．

当已知F为正态时，习用的是t区间估计，对一般情况，因θ为总体中位数，可用§2·4的三段中的方法去处理．但那个方法没有用到F为对称分布这个特点，现在介绍一种方法，是以这一点为依据．注意：在以下总假定$F(x)$关于0对称且处处连续，不再一一申明．

这个方法分以下几个步骤：

1．找一个统计量 $T=T(X_1,\cdots,X_n)$，使：

（1）当 $\theta=0$ 时，T 的分布不依赖于 F（即统计量 T 在原假设 $\theta=0$ 时分布无关）。

（2）对任何实数 $x_1,\cdots,x_n$，$T(x_1-\theta,\cdots,x_n-\theta)$ 作为 θ 的函数，是非增的。

在以下用 P_θ 表示：事件的概率是在参数值为 θ 时计算的。

2．找 d_1,d_2，使

$$P_0(T<d_1)=\frac{\alpha}{2},\ P_0(T>d_2)=\frac{\alpha}{2}, \tag{4·171}$$

此处 $1-\alpha$ 为预定的置信系数。

3．根据1(2)，可找到统计量 $\hat{\theta}_i=\hat{\theta}_i(X_1,\cdots,X_n)$，$i=1,2$，使

$$T(X_1-\theta,\cdots,X_n-\theta)\leqslant d_2\Longleftrightarrow\theta\geqslant\hat{\theta}_1, \tag{4·172}$$

$$T(X_1-\theta,\cdots,X_n-\theta)\geqslant d_1\Longleftrightarrow\theta\leqslant\hat{\theta}_2, \tag{4·173}$$

则$[\hat{\theta}_1,\ \hat{\theta}_2]$就是 θ 的置信系数 $1-\alpha$ 的区间估计。

这个方法的主要之点当然在于1(1)。方法本身并不直接与秩统计量相关联。但是，由于满足条件1(1)的统计量多是秩统计量，说此方法本质上是一种秩方法，亦未尝不可。

方法的证明很简单：设 θ 的真值为 θ_0。由(4·172)和(4·173)，有

$$\begin{aligned}P_{\theta_0}(\hat{\theta}_1\leqslant\theta_0\leqslant\hat{\theta}_2)&=P_{\theta_0}(d_1\leqslant T(X_1-\theta_0,\cdots,X_n-\theta_0)\leqslant d_2)\\&=P_0(d_1\leqslant T(X_1,\cdots,X_n)\leqslant d_2)\end{aligned} \tag{4·174}$$

再用(4·171)，即得 $P_{\theta_0}(\hat{\theta}_1\leqslant\theta_0\leqslant\hat{\theta}_2)=1-\alpha$。这证明了$[\hat{\theta}_1,\ \hat{\theta}_2]$确有置信系数$1-\alpha$。

在以上的叙述中忽略了某些细节：

a．由于秩统计量只取有限个值，当使用秩统计量 T 于以上方法时，不一定能找到 d_1，d_2，使严格地满足(4·171)。这时，只能或者修改 α 之值，或使用随机化，后者在应用上是尽力避免的，故只有修改 α 之一途，但如 n 太小，则 α 调整量可能过大，而不适合问题的要求。因此，本质上说，这方法虽不基于大样本

性质，但 n 仍不能太小。

b. 有时，（4·172）及（4·173）中的"$\Longleftrightarrow$"号的右边，可能不是 $\theta \geqslant \hat{\theta}_1$ 或 $\theta \leqslant \hat{\theta}_2$，而是 $\theta > \hat{\theta}_1$ 或 $\theta < \hat{\theta}_2$。与此相应，（4·174）的第一项，也要修改为 $P_{\theta_0}(\hat{\theta}_1 < \theta_0 < \hat{\theta}_2)$，或 $P_{\theta_0}(\hat{\theta}_1 < \theta_0 \leqslant \hat{\theta}_2)$，或 $P_{\theta_0}(\hat{\theta}_1 \leqslant \theta_0 < \hat{\theta}_2)$。因而得到的区间估计是四种可能（区间左、右端开、闭的四种组合）之一。这在应用上并无重要性，但值得注意一下。

c. 当 n 较大时，直接从 $T(X_1, \cdots, X_n)$ 在 $\theta=0$ 时的分布出发去找 d_1, d_2 太繁，可使用大样本逼近，这只要求 $T(X_1, \cdots, X_n)$ 在 $\theta=0$ 之下有极限分布就可以了，因为（4·171）式只涉及 $\theta=0$ 时的分布。

例4·13 取 Wilcoxon 符号秩和统计量 $W^+ = W^+(X_1, \cdots, X_n)$ 作为统计量 T，我们来验证，它符合1(1)和1(2)。前者由定理 4·9 直接得出，只须注意：定理 4·9 中关于符号秩的分布的(1)～(3)几条中，没有一条涉及分布 F。1(2)可由公式（4·69）推出。据该公式，有

$$W^+(X_1, \cdots, X_n) = \sum_{1 \leqslant i \leqslant j \leqslant n} \psi(X_i + X_j) = \sum_{1 \leqslant i \leqslant j \leqslant n} \psi\left(\frac{X_i + X_j}{2}\right),$$

其中 $\psi(x) = I(x > 0)$。于是有

$$W^+(x_1 - \theta, \cdots, x_n - \theta) = \sum_{1 \leqslant i \leqslant j \leqslant n} \psi\left(\frac{x_i + x_j}{2} - \theta\right),$$

由于 $\psi(x)$ 是 x 的非降函数，知1(2)成立。

W^+ 只取 $0, 1, 2, \cdots, n(n+1)/2$ 等数为值，且易见当 $\theta=0$ 时，W^+ 的分布关于 $n(n+1)/4$ 点对称（换句话说，有 $P_0(W^+ = i) = P_0\left(W^+ = \frac{n(n+1)}{2} - i\right), i = 0, 1, \cdots, \frac{n(n+1)}{2}$。这个简单事实的证明留给读者（习题18）。于是先找出 d_1，使

$$\sum_{i=0}^{d_1 - 1} P_0(W^+ = i) = \alpha/2, \qquad (4·175)$$

然后取 $d_2 = \frac{n(n+1)}{2} - d_1$ 即可。

最后，要利用定出的 d_1，d_2，根据（4·172）和（4·173）定出 $\hat{\theta}_1$ 和 $\hat{\theta}_2$。为此，把 $N=\frac{1}{2}n(n+1)$ 个数 $\left\{\frac{X_i+X_j}{2}: 1\leqslant i\leqslant j\leqslant n\right\}$ 按由小到大排列为 $W_{(1)}\leqslant W_{(2)}\leqslant\cdots\leqslant W_{(N)}$. 则易见

$$W^+(X_1-\theta,\cdots,X_n-\theta)\leqslant d_2 \Longleftrightarrow \theta_2\geqslant W_{(d_1)}, \qquad (4\cdot176)$$

$$W^+(X_1-\theta,\cdots,X_n-\theta)\geqslant d_1 \Longleftrightarrow \theta_1\leqslant W_{(d_2)}. \qquad (4\cdot177)$$

事实上，若 $W^+(X_1-\theta,\cdots,X_n-\theta)\leqslant d_2$，则在$N$个值 $\left\{\frac{X_i+X_j}{2}-\theta: 1\leqslant i\leqslant j\leqslant n\right\}$中，至多有 d_2 个大于 0. 即在N 个值$\left\{\frac{X_i+X_j}{2}: 1\leqslant i\leqslant j\leqslant n\right\}$中，至多有 d_2 个大于 θ。由于 $d_1+d_2=N$，这无异乎说 $\theta\geqslant W_{(d_1)}$。同样的理由得出，当 $\theta\geqslant W_{(d_1)}$ 时有 $W^+(X_1-\theta,\cdots,X_n-\theta)\leqslant d_2$、于是证明了（4·176），（4·177）的证明类似。

这样，得到 θ 的置信系数 $1-\alpha$ 的区间估计为

$$[W_{(d_1)},\ W_{(d_2)}] \qquad (4\cdot178)$$

如果（4·175）式只是近似的，则 $1-\alpha$ 也只是近似的置信系数。

当 n 较大时，可利用 W^+在 $\theta=0$ 之下的极限分布（见例4·5），而近似地定出

$$d_1=\frac{n(n+1)}{4}-\left(\frac{n(n+1)(2n+1)}{24}\right)^{1/2}u_{\alpha/2},$$

$$d_2=\frac{n(n+1)}{4}+\left(\frac{n(n+1)(2n+1)}{24}\right)^{1/2}u_{\alpha/2}.$$

这样定出的 d_1，d_2 一般不是整数，可以把 d_1 修正为 $[d_1]$（不超过 d_1 的最大整数），而 d_2 修正为$\frac{n(n+1)}{2}-[d_1]$。

这个方法给人印象深刻之处在于理论简单，而在很广的分布族之下，得出有确切置信系数的区间估计。人们往往在并不知道总体分布是否为正态时也使用 t 区间估计，而求助于这个事实：当样本大小 $n\to\infty$ 时，其置信系数趋于 $1-\alpha$。然而，对固定的 n（不管多大），总可以找到连续的对称分布 F，使对这个 F 而言，

t 区间的置信系数小于指定的 $\varepsilon>0$。所以，如果我们对问题中的总体分布确实了解很少，则使用 t 区间估计可能带来大的失误，而用（4·178）这类估计，就没有这个问题。

还有一个效率的问题。比方说，总体分布 F 确为正态。这时，（4·178）和 t 区间估计都可用，二者的相对效率如何？这就需要引进一种与假设检验情况类似的渐近相对效率 ARE。我们不去深入其细节，而只指出其结论是：区间估计之间的 ARE，正好就是其所相应的检验之间的 ARE。拿本例情况说，当分布真为正态时，（4·178）对 t 区间估计的 ARE，就是 Wilcoxon 符号秩和检验对一样本 t 检验的 ARE，即 $3/\pi$。如果总体分布为其他分布，则 ARE 相应地变化。从例 4·5 的讨论可知，（4·178）对 t 区间估计处在有利地位。

所有这一切使我们相信：相对于优良性能来说，非参数方法在目前实用中受到的注意太小，而基于正态假定的方法的状况，则正好反过来。这原因我们在第一章中已有所说明，有历史的、计算的（非参数方法往往涉及较繁复的计算，如果不使用极限分布的话）等原因。一部分也是出于误解，认为非参数方法貌似粗糙，可能比具有精细理论作背景的正态方法，在效率上要差得多。我们相信，这种状况有朝一日会起一定的变化，也是很可能的。

二、位置参数的区间估计

设 $X_1,\cdots,X_{n_1}$ 和 $Y_1,\cdots,Y_{n_2}$ 分别是从分布 $F(x)$ 及 $F(x-\theta)$ 中抽出的简单样本，要作 θ 的区间估计。此处我们只假定 F 处处连续（不必对称），其余全未知。

解法步骤与对称中心的情况完全类似。

1. 找一个统计量 $T=T(X_1,\cdots,X_{n_1},Y_1,\cdots,Y_{n_2})$，使：

(1) 当 $\theta=0$ 时，T 的分布不依赖于 F；

(2) 对任何实数 $x_1,\cdots,x_{n_1},y_1,\cdots,y_{n_2}$，$T(x_1,\cdots,x_{n_1},y_1-\theta,\cdots,y_{n_2}-\theta)$ 作为 θ 的函数，是非增的。

以下 P_θ 的意义与前面相同。

2. 找 d_1, d_2，使(4·171)成立，其中 $1-\alpha$ 是预定的置信系数。

3. 据1(2)，找统计量 $\hat{\theta}_i=\hat{\theta}_i(X_1,\cdots,X_{n_1},Y_1,\cdots,Y_{n_2})$，$i=1$，2，使

$$T(X_1,\cdots,X_{n_1},\ Y_1-\theta_1,\cdots,Y_{n_2}-\theta_1)\leqslant d_2 \Longleftrightarrow \theta\geqslant\hat{\theta}_1,$$

$$T(X_1,\cdots,X_{n_1},Y_1-\theta_2,\cdots,Y_{n_2}-\theta_2)\geqslant d_1 \Longleftrightarrow \theta\leqslant\hat{\theta}_2,$$

则$[\hat{\theta}_1,\hat{\theta}_2]$为 θ 的一个置信系数 $1-\alpha$ 的区间估计。证明与对称中心的情况完全相似，细节留给读者。又在对称中心情况下讲的 a 、 b 、 c 三条注意 见(4·174)式后面)，在此也完全适用。

例4·13 选用 Wilcoxon 秩和统计量 $W=R_1+\cdots+R_{n_2}$，其中 R_i 是 Y_i 在合样本 $X_1\ \cdots,X_{n_1},\ Y_1,\cdots,Y_{n_2}$ 中之秩。当 $\theta=0$ 时，合样本独立同分布，有公共分布 F。因 F 连续，由定理 4·1 知 W 的分布与 F 无关，故条件 1(1)满足。条件 1(2) 易验证(细节留给读者)。

为了找 d_1,d_2，利用公式(3·10)，将 W 表为 $W=U+n_2(n_2+1)/2$，其中 U 为集合

$$A=\{Y_j-X_i:1\leqslant i\leqslant n_1,1\leqslant j\leqslant n_2\} \qquad (4·179)$$

中大于 0 的个数。据定理 4·2，当 $\theta=0$ 时，W 的分布关于 $n_2(n+1)/2$ 点对称，$n=n_1+n_2$。故 U 之分布关于点 $n_2(n+1)/2-n_2(n_2+1)/2=n_1n_2/2$ 对称。以下就用 U 代替 W 来讨论。取 d_1，使

$$\sum_{i=0}^{d_1-1}P_0(U=i)=\frac{\alpha}{2}, \qquad (4·180)$$

再取 $d_2=n_1n_2-d_1$ 即可。最后要利用 d_1,d_2 定出 $\hat{\theta}_1$ 和 $\hat{\theta}_2$。其推理过程与得出(4·176)和(4·177)者相似。我们只把结果写出如下：把(4·179)的集 A 中各元按由小到大排列为 $U_{(1)}\leqslant\cdots\leqslant U_{(n_1 n_2)}$，则 $\hat{\theta}_1=U_{(d_1)}$ 而 $\hat{\theta}_2=U_{(d_2+1)}$，从而得出 θ 的区间估计

$$[U_{(d_1)},U_{(d_2+1)}] \qquad (4·181)$$

只要(4·180)是确切的，则区间估计(4·181)有确切的置信系数 $1-\alpha$，不管总体分布 F 如何。

这个区间估计对两样本 t 区间估计的 ARE，正好等于

wilcoxon 秩和检验对两样本 t 检验的 ARE。例如，对正态模型这个最有利于 t 区间估计的场合，ARE 为 $3/\pi$。

我们也可以选定一个计分函数 $a(\cdot)$，而从统计量 $L_n=\sum_{i=1}^{n_2} a(R_i)$ 出发，去构造区间估计。这区间估计对 t 区间估计的 ARE，正好等于基于 L_n 的两样本秩检验对两样本 t 检验的 ARE。例如，若取 $a(i)=\Phi^{-1}\left(\frac{i}{n+1}\right)$ 或 $E(\Phi^{-1}(U_{ni}))$，其中 Φ^{-1} 为 $N(0,1)$的分布函数的反函数，而 $U_{n1}\leqslant\cdots\leqslant U_{nn}$ 为$(0,1)$均匀分布的次序样本，则所产生的区间估计，其对 t 区间估计的 ARE，在正态模型下正好为 1，而在其他模型下总不小于 1。

三、对称中心的点估计

给对称中心作点估计的方法，以前提了好几种，像样本均值（当总体期望存在时），样本中位数、截尾均值、Winsor 化均值等。这些估计在一般情形下也够用了。那么这个问题是否还值得考虑。我们说有这样一个理由。不同的估计，相对于不同的总体分布而言，其优越性各不同。如果我们对一个量设计了多种估计法，而又了解各个估计在何种模型下性能较优，则当我们对所面对的问题的模型有所了解时，可以从这些估计中选用一个，其性能较为优良。

这里我们介绍 Hodges 和 Lehmann 在 1963 年引进的一种估计法。与前面介绍的区间估计法相似，这方法形式上并不一定与秩有关，但其使用多半限于秩统计量的情形。

现设 $X_1,\cdots,X_n$ 为抽自分布 $F(x-\theta)$ 的简单样本，F 关于0对称且处处连续，其他无所知。要据此估计 θ。Hodges-Lehmann 估计的步骤如下：

1．找一统计量 $T=T(X_1,\cdots,X_n)$，具有以下的性质：

(1) 当 $\theta=0$ 时，T的分布关于某点 c 对称。c 为已知常数，与F无关。

(2) 对任何实数 $x_1,\cdots,x_n$，$T(x_1+\theta,\cdots,x_n+\theta)$作为 θ 的函

数，是非降的。

2. 定义 $\hat{\theta}_i=\hat{\theta}_i(X_1,\cdots,X_n)$, $i=1,2$, 如下：

$$\hat{\theta}_1=\sup\{a:T(X_1-a,\cdots,X_n-a)>c\}, \tag{4·182}$$

$$\hat{\theta}_2=\inf\{a:T(X_1-a,\cdots,X_n-a)<c\}, \tag{4·183}$$

条件 1 (2)保证了 $\hat{\theta}_1\leqslant\hat{\theta}_2$。

3. 用 $\hat{\theta}=\hat{\theta}(X_1,\cdots,X_n)=\frac{1}{2}(\hat{\theta}_1+\hat{\theta}_2)$作为 θ 的估计。这就是 Hodges-Lehmann 的估计，以下简称为 HL 估计。

所以，HL 估计不是一个固定的估计。随着统计量T的选择的不同，可以得到种种不同的估计。如下面将看到的，其优良性取决于总体分布F如何。

本方法的关键之点在于选择统计量T。T的选择，一般是使用为检验原假设 $\theta=0$ 时的检验统计量。值得注意的是，此处并不要求在 $\theta=0$ 时，T 的分布与F无关。故不一定要从非参数检验统计量出发，见下例。

例4·14 若使用一样本 t 统计量

$$T=\sqrt{n}\,\overline{X}/S,\quad \overline{X}=\sum_{i=1}^{n}X_i/n,\quad S^2=\sum_{i=1}^{n}(X_i-\overline{X})^2/(n-1)$$

则易见条件 $1a$ 和 $1b$ 都满足(验证细节留给读者)，且 $c=0$，$\hat{\theta}_1=\hat{\theta}_2=\overline{X}$。由此得出估计量 $\overline{X}$。当总体分布F为正态时，这估计量$\overline{X}$有很多优越性，而 t 统计量又正好是当F为正态时，检验 $\theta=0$ 的优良方法。以下将指出这不是巧合。

如果选用符号统计量

$$T=\sum_{i=1}^{n}I(X_i>0),$$

则易见条件 $1a$ 和 $1b$ 都满足，$c=n/2$，当 n 为奇数时，$\hat{\theta}_1=\hat{\theta}_2=X_{(\frac{n+1}{2})}$。当 n 为偶数时，有 $\hat{\theta}_1=X_{(n/2)}$，$\hat{\theta}_2=X_{(n/2+1)}$ 由这两种情况都得到 $\hat{\theta}=\text{med}(X_1,\cdots,X_n)$(参看(2·17)式)。

对 HL 估计不难证明下面的小样本性质：

1. 估计 $\hat{\theta}$ 有“平移同变性”，即对任何常数 a，有

$$\hat{\theta}(X_1+a,\cdots,X_n+a)=\hat{\theta}(X_1,\cdots,X_n)+a \tag{4·184}$$

为证此，只须验证对 $i=1, 2$，有 $\hat{\theta}_i(X_1+a, \cdots, X_n+a)=\hat{\theta}_i(X_1,\cdots,X_n)+a$。这不难从 $\hat{\theta}_i$ 的定义得到。细节留给读者。

2. 若在定义 HL 估计时所选定的统计量T，除满足条件 1a 和 1b 外，还满足

$$T(-x_1,\cdots,-x_n)=2c-T(x_1,\cdots,x_n) \qquad (4\cdot185)$$

其中 c 就是条件(1)中提到的那个 c 。则 $\hat{\theta}$ 的分布关于 θ 点对称，且

$$\hat{\theta}(-x_1,\cdots,-x_n)=-\hat{\theta}(x_1,\cdots,x_n) \qquad (4\cdot186)$$

为证此，利用 $\hat{\theta}_1$ 的定义及（4·185）式，得

$$\begin{aligned}\hat{\theta}_1(-x_1,\cdots,-x_n)&=\sup\{a:T(-x_1-a,\cdots,-x_n-a)>c\}\\&=\sup\{a:(2c-T(x_1+a,\cdots,x_n+a))>c\}\\&=\sup\{a:T(x_1+a,\cdots,x_n+a)<c\},\end{aligned}$$

在此式中，改 a 为$-a$，并注意 $\sup\{-a:a\in A\}=-\inf\{a:a\in A\}$，得

$$\begin{aligned}\hat{\theta}_1(-x_1,\cdots,-x_n)&=-\inf\{a:T(x_1-a,\cdots,x_n-a)<c\}\\&=-\hat{\theta}_2(x_1,\cdots,x_n),\end{aligned}$$

在此式中以 $-x_i$ 代替 x_i，又得

$$\hat{\theta}_2(-x_1,\cdots,-x_n)=-\hat{\theta}_1(x_1,\cdots,x_n).$$

由以上两式即得（4·186）。又因 X_i 之分布关于 θ 对称，有 $X_i\overset{d}{=}2\theta-X_1$，于是由（4·184）及（4·186），得

$$\begin{aligned}\hat{\theta}(X_1,\cdots,X_n)&\overset{d}{=}\hat{\theta}(2\theta-X_1,\cdots,2\theta-X_n)\\&=2\theta+\hat{\theta}(-X_1,\cdots,-X_n)\\&=2\theta-\hat{\theta}(X_1,\cdots,X_n),\end{aligned}$$

这证明了 $\hat{\theta}$ 之分布关于 θ 对称。由此可知，$\text{med}(\hat{\theta})=\theta$。在统计上，称具有这个性质的估计量 $\hat{\theta}$ 为"中位无偏"的。若 $E|\hat{\theta}|<\infty$，则也有 $E_\theta(\hat{\theta})=\theta$，即 $\hat{\theta}$ 为 θ 的通常意义下的无偏估计。但$\hat{\theta}$的期望是否存在有限，取决于总体分布F及所选定的统计量T。如在例 4·13 中，所得估计量$\bar{X}$是否有期望，要看总体分布F是否有期望。另一估计量，$X_1,\cdots,X_n$ 的样本中位数，它的期望存在的条件则较此为低。

关于 HL 估计的大样本性质，最重要的是下面的结果：

定理4·13 以 $X_1,\cdots,X_n$ 记从分布 $F(x-\theta)$ 中抽出的简单样本，$F(x)$关于0对称且处处连续。对每个自然数 n 选定统计量 $T_n=T_n(X_1,\cdots,X_n)$，适合为构造 θ 的 HL 的条件 1(1) 和 1(2)，条件1(1)中的常数 c 现记为 c_n，所产生的 HL 估计记为 $\hat{\theta}_n$。如果统计量 T_n 满足定理 4·11 的条件（即条件（4·54）、（4·56）、（4·58）、（4·60）和（4·62），且 $c_n=\mu(T,n,o,F)$与 F 无关，则当 $n\to\infty$时有

$$\sqrt{n}\ (\hat{\theta}_n-\theta)\xrightarrow{\mathscr{L}}N(0,(K_T^2(F))^{-1}).\qquad(4\cdot187)$$

证明 把在定义 HL 估计 $\hat{\theta}_n$ 的过程中产生的统计量 $\hat{\theta}_1$ 和 $\hat{\theta}_2$ 记为 $\hat{\theta}_{1n}$ 和 $\hat{\theta}_{2n}$。由（4·184）式有 $\hat{\theta}_n(X_1,\cdots,X_n)-\theta=\hat{\theta}_n(X_1-\theta,\cdots X_n-\theta)$，且因 $X_i\sim F(x-\theta)$有 $X_i-\theta\sim F(x)$，故不失普遍性可设 $\theta=0$，固定常数 a，由（4·182）和（4·183），有

$$T_n\left(X_1-\frac{a}{\sqrt{n}},\cdots,X_n-\frac{a}{\sqrt{n}}\right)>c_n$$

$$\Rightarrow\begin{cases}\hat{\theta}_{1n}(X_1,\cdots,X_n)\geqslant a/\sqrt{n}\\ \hat{\theta}_{2n}(X_1,\cdots,X_n)\geqslant a/\sqrt{n}\end{cases}\Rightarrow\hat{\theta}_n(X_1,\cdots,X_n)\geqslant a/\sqrt{n}$$

于是得到

$$\begin{aligned}&P_0(\sqrt{n}\,\hat{\theta}_n(X_1,\cdots,X_n)\geqslant a)\\&\geqslant P_0(T_n(X_1-\frac{a}{\sqrt{n}},\cdots,X_n-\frac{a}{\sqrt{n}})>c_n)\\&=P_{-a/\sqrt{n}}(T_n(X_1,\cdots,X_n)>c_n)\\&=P_{-\frac{a}{\sqrt{n}}}\left(\frac{T_n-\mu\left(T,n,\frac{-a}{\sqrt{n}},F\right)}{\sigma\left(T,n,\frac{-a}{\sqrt{n}},F\right)}\right.\\&\left.>\frac{\mu(T,n,o,F)-\mu\left(T,n,\frac{-a}{\sqrt{n}},F\right)}{\sigma\left(T,n,\frac{a}{\sqrt{n}},F\right)}\right)\end{aligned}\qquad(4\cdot188)$$

根据假定，T_n 满足（4，54）。故当 $n\to\infty$ 时，变量 $\left(T_n-\mu\left(T,n,\frac{a}{\sqrt{n}},F\right)\right)/\sigma\left(T,n,\frac{a}{\sqrt{n}},F\right)$ 在 $\theta=-\frac{a}{\sqrt{n}}$ 之下的分布，收敛于标准正态分布 Φ。再据 μ 及 σ 满足（4·58）、（4·60）和（4·62），知当 $n\to\infty$ 时，有

$$\left(\mu(T,n,o,F)-\mu\left(T,n,\frac{-a}{\sqrt{n}},F\right)\right)/\sigma\left(T,n,\frac{a}{\sqrt{n}},F\right)\longrightarrow aK_T(F),$$

于是由（4·188），得

$$\liminf_{n\to\infty}P_0\,(\sqrt{n}\,\theta_n\geqslant a)\geqslant 1-\Phi(aK_T(F))\tag{4·189}$$

另一方面，仍据（4·182）和（4·183），有

$$T_n\left(X_1-\frac{a}{\sqrt{n}},\cdots,X_n-\frac{a}{\sqrt{n}}\right)<c_n$$

$$\Rightarrow\left\{\begin{aligned}&\hat{\theta}_{1n}(X_1,\cdots,X_n)\leqslant\frac{a}{\sqrt{n}}\\&\hat{\theta}_{2n}(X_1,\cdots,X_n)\leqslant\frac{a}{\sqrt{n}}\end{aligned}\right\}\Rightarrow\hat{\theta}_n(X_1,\cdots,X_n)\leqslant\frac{a}{\sqrt{n}},$$

重复上面的推理方法，又可得到

$$\liminf_{n\to\infty}P_0\,(\sqrt{n}\,\theta_n\leqslant a)\geqslant\Phi(aK_T(F)).\tag{4·190}$$

注意到（4·189）的左边等于 $1-\limsup\limits_{n\to\infty}P_0(\sqrt{n}\,\theta_n<a)$，得

$$\limsup_{n\to\infty}P_0(\sqrt{n}\,\hat{\theta}_n<a)\leqslant\Phi(aK_T(F)),\tag{4·191}$$

由（4·190）和（4·191），并注意到分布函数 Φ 处处连续，即得

$$\lim_{n\to\infty}P_0(\sqrt{n}\,\theta_n\leqslant a)=\Phi(aK_T(F)).$$

于是证明了（4·187）。

如果把两个估计量的渐近方差倒数之比作为其渐近相对效率，则定理4·13可以解释为：设我们从两个统计量 S_n 和 T_n 出发分别去作 θ 的 HL 估计，结果记为 $\hat{\theta}_{nS}$ 和 $\hat{\theta}_{nT}$，而 S_n 和 T_n 都满足定理4·13的条件，则 $\hat{\theta}_{nS}$ 对 $\hat{\theta}_{nT}$ 的渐近相对效率，记为

$\mathrm{ARE}(\hat{\theta}_S,\hat{\theta}_T;\ F)$，等于 $K_S^2(F)/K_T^2(F)$，即等于用 S 和 T 去检验假设 $\theta=0$ 时的渐近相对效率 $\mathrm{ARE}(S,T;\ F)$。于是前面讲过的有关 $\mathrm{ARE}(S,T;\ F)$ 的一切，都可移于此处。这个定理也印证了本段开头处所阐述的那个论点：即我们可使用非参数性的方法构造出 θ 的一些点估计，它们针对不同的总体分布 F 各有其优越性。一旦我们对 F 有所了解，就可据以选定一个适当的估计。

总结前面有关检验、区间估计和点估计的讨论看出：所有关于渐近相对效率的定义，最后都归结为 Pitman 的 ARE。由此我们相信：Pitman 的 ARE 确实抓着了大样本效率的实质所在。

四、位置参数的点估计

设 $X_1,\cdots,X_{n_1}$ 和 $Y_1,\cdots,Y_{n_2}$ 是从连续分布 $F(x)$ 和 $F(x-\theta)$ 中抽出的简单样本，要估计 θ。用 HL 方法估计 θ 的步骤及所得估计量的性质，与对称中心估计的场合完全相似。故以下只把有关步骤和结论列出，建议读者自己把所有细节补出来。

1．找统计量 $T=T(X_1,\cdots,X_{n_1},\ Y_1,\cdots,Y_{n_2})$，满足条件：

（1）当 $\theta=0$ 时，T 的分布关于某点 c 对称，c 已知且与 F 无关。

（2）对任何实数 $x_1,\cdots,x_{n_1},y_1,\cdots,y_{n_2}$，$T(x_1,\cdots,x_{n_1},\ y_1+\theta,\cdots,y_{n_2}+\theta)$ 作为 θ 的函数，是非降的。

2．定义 $\hat{\theta}_i=\hat{\theta}_i(X_1,\cdots,X_{n_1},\ Y_1,\cdots,Y_{n_2})$，$i=1,2$，如下：

$\hat{\theta}_1=\sup\{a:T(X_1,\cdots,X_{n_1},\ Y_1-a,\cdots,Y_{n_2}-a)>c\}$，

$\hat{\theta}_2=\inf\{a:T(X_1,\cdots,X_{n_1},Y_1-a,\cdots,Y_{n_2}-a)<c\}$。

3．用 $\hat{\theta}(X_1,\cdots,X_{n_1},Y_1,\cdots,Y_{n_2})=\hat{\theta}=\frac{1}{2}(\hat{\theta}_1+\hat{\theta}_2)$ 估计 θ。

例如，取 $T=$Wilcoxon 秩和统计量，则易见条件1(1)和1(2)都满足，且可算出

$$\hat{\theta}=\mathrm{med}\{Y_j-X_i:\ i=1,\cdots,n_1,\ j=1,\cdots,n_2\},\qquad (4\cdot192)$$

上式右边是 n_1n_2 个数 $\{Y_j-X_i\}$ 的样本中位数。（4·192）的证明留给读者。

上面定义的 HL 估计有性质：

1. 平移同变性：对任何常数 a 有

$$\hat{\theta}(X_1,\cdots,X_{n_1},Y_1+a,\cdots,Y_{n_2}+a)$$
$$=\hat{\theta}(X_1,\cdots,X_{n_1},\ Y_1,\cdots,Y_{n_2})+a$$

2. 若统计量 T 除满足条件 1(1)、1(2)外，还满足：

1° 对任何 $x_1,\cdots,x_{n_1},\ y_1,\cdots,y_{n_2}$ 有

$$T(-x_1,\cdots,-x_{n_1},\ -y_1,\cdots,-y_{n_2})$$
$$=2c-T(x_1,\cdots,x_{n_1},\ y_1,\cdots,y_{n_2})$$

2° 对任何常数 a 有

$$T(x_1+a,\cdots,x_{n_1}+a,\ y_1+a,\cdots,y_{n_2}+a)$$
$$=T(x_1,\cdots,x_{n_1},y_1,\cdots,y_{n_2}),$$

又分布函数 $F(x)$ 关于某点对称，则 $\hat{\theta}$ 的分布关于 θ 对称，且

$$\hat{\theta}(-x_1,\cdots,-x_{n_1},-y_1,\cdots,-y_{n_2})=-\hat{\theta}(x_1,\cdots,x_{n_1},\ y_1,\cdots,y_{n_2})$$

关于 $\hat{\theta}$ 的大样本性质，成立着与定理 4·13 类似的定理。

定理4·14 记 $n=n_1+n_2$。把 θ 的 HL 估计记为 $\hat{\theta}_n$，又定义中涉及的常数记为 c_n。若统计量 T_n 满足定理 4·11 的条件，且 $n_1/n\to\lambda$ 当 $n\to\infty$，$0<\lambda<1$。又 $c_n=\mu(T,n,0,F)$，与 F 无关。则当 $n\to\infty$ 时有

$$\sqrt{n}\ (\hat{\theta}_n-\theta)\xrightarrow{\mathscr{L}}N(0,(\lambda(1-\lambda)K_T^2(F))^{-1}). \qquad (4·193)$$

根据这个定理，若以渐近方差倒数之比来衡量两估计量间的渐近相对效率，则两个 HL 估计之间的渐近相对效率，正好等于用这两个统计量所作的对 $\theta=0$ 的检验的渐近相对效率。

§4.6 Смирнов检验与 Колмогоров检验

Смирнов 检验在本书开篇的例 1·1 中就提到过了。不难看出，它只涉及样本的秩，因而本质上是一个秩检验，放到本章讨论可以说得通。至于 Колмогоров 检验，则不是一个秩检验，放

到本章，于体例不合。可是我们不好为这个检验设专章，且它与 Смирнов 检验，在使用经验分布这一点上，有其共通之处，故也放在这一节一并讨论。虽于体例有所不合，也顾不得了。

一、Смирнов 检验

设 $X_1,\cdots,X_{n_1}$ 和 $Y_1,\cdots,Y_{n_2}$ 分别是从分布 F 和 G 中抽出的简单样本，要检验假设 $F=G$。以 F_{n_1} 和 G_{n_2} 分别记 $X_1,\cdots,X_{n_1}$ 和 $Y_1,\cdots,Y_{n_2}$ 的经验分布函数，其定义见（1·4）和（1·5）（改 m 为 n_1，n 为 n_2）。由于 F_{n_1} 和 G_{n_2} 分别是 F 和 G 的估计，知当原假设 $F=G$ 成立时，F_{n_1} 和 G_{n_2} 应接近。因此，若令

$$S_{n_1n_2}=\sup_{-\infty<x<\infty}|F_{n_1}(x)-G_{n_2}(x)| \tag{4·194}$$

则当原假设成立时，$S_{n_1n_2}$ 应倾向于小。故以

$$\{S_{n_1n_2}>C\} \tag{4·195}$$

为否定域的检验，是"$F=G$"的一个可用的检验。

为了根据给定的检验水平 α 确定临界值 C，需要定出在原假设成立时，$S_{n_1n_2}$ 的分布。这个在原则上不难。因为易见，在 F、G 都连续，因而以概率 1 在合样本 $X_1,\cdots,X_{n_1},Y_1,\cdots,Y_{n_2}$ 中无结存在时，统计量 $S_{n_1n_2}$ 只与 $Y_1,\cdots,Y_{n_2}$ 在合样本中之秩 $R_1,\cdots,R_{n_2}$ 有关（这一简单事实的证明留给读者），因而可利用定理 4·1 定出 $S_{n_1n_2}$ 的分布。例如，在 $n_1=3$，$n_2=2$ 的场合，(R_1,R_2) 以等概率 $\left(各\frac{1}{20}\right)$ 取以下 20 组值之一：

(1,2)，(1,3)，(1,4)，(1,5)，(2,3)(2,4)，(2,5)，(3,4)，(3,5)，(4,5)，(2,1)，(3,1)，(4,1)，(5,1)，(3,2)，(4,2)，(5,2)，(4,3)，(5,3)，(5,4)

与之相应的 S_{32} 之值分别为：

$1,\ \frac{2}{3},\ \frac{1}{2},\ \frac{1}{2},\ \frac{2}{3},\ \frac{1}{3},\ \frac{1}{2},\ \frac{2}{3},\ \frac{2}{3},\ 1,$

$1,\ \frac{2}{3},\ \frac{1}{2},\ \frac{1}{2},\ \frac{2}{3},\ \frac{1}{3},\ \frac{1}{2},\ \frac{2}{3},\ \frac{2}{3},\ 1$

由此得出在原假设$F=G$成立且F、G处处连续时，S_{32}的概率分布为

$$P(S_{32}=1)=\frac{1}{5},\ P\left(S_{32}=\frac{2}{3}\right)=\frac{2}{5},\ P\left(S_{32}=\frac{1}{2}\right)=\frac{3}{10},$$
$$P\left(S_{32}=\frac{1}{3}\right)=\frac{1}{10}$$

对一般的n_1n_2，原则上没有什么困难。定出$S_{n_1n_2}$在原假设下的分布后，临界值C即可根据给定的显著性水平α定出。往往对某个特定的α（如$\alpha=0.05$）不存在常数C，使当$F=G$时恰有$P(S_{n_1n_2}>C)=\alpha$。这时，或者适当调整α之值，或者实行随机化。Harter 和 Owen 的表《Selected Tables in Mathematical Statistics》，Vol. 3 载有n_1和n_2都不超过100的情况，也可参看数学所概统室编的《常用数理统计表》。

在n_1和n_2都很大时，可使用 Смирнов 在1939年证明的下述极限定理：

定理4·15 当$F=G$连续，且存在$\lambda>0$使当$n_1\to\infty$，$n_2\to\infty$时，始终保持$n_2/n_1<\lambda$和$n_1/n_2<\lambda$，则当$n_1\to\infty$，$n_2\to\infty$时，有

$$\sqrt{\frac{n_1n_2}{n_1+n_2}}S_{n_1n_2}\xrightarrow{\mathscr{L}}K(x),\tag{4·196}$$

其中

$$K(x)=\begin{cases}0, & \text{当 } x\leqslant 0;\\ \sum_{i=-\infty}^{\infty}(-1)^i\exp(-2i^2x^2), & \text{当 } x>0.\end{cases}\tag{4·197}$$

这定理证明的方法有好几种，但都较繁，无法在这里细讲。分布$K(x)$的95%和99%分位点分别为1.358和1.628，故当α取为0.05或0.01时，C之值（当n_1，n_2大时）可近似地取为$1.358\sqrt{\frac{n_1+n_2}{n_1n_2}}$或者$1.628\sqrt{\frac{n_1+n_2}{n_1n_2}}$。

Смирнов 检验（4·194）和（4·195）是"面向四方"的，即一切可能的对立假设，都在其考虑之列。因此，除非在事先对可能

的对立假设确是了解很少，一般不大使用这检验，因为其效率恐不及在前几节中讨论过的那种更有针对性的检验。但如我们事先知道可能的对立假设是

$$Y \overset{r}{>} X\text{，或 } G(x) \leqslant F(x) \text{对一切 } x\text{，但 } F \neq G \tag{4·198}$$

则

$$S_{n_1 n_2}^{+} = \sup_{-\infty < x < \infty} (F_{n_1}(x) - G_{n_2}(x)) \tag{4·199}$$

是一个合适的统计量。有趣的是：在原假设$F = G$成立之下，$S_{n_1 n_2}^{+}$的极限定理的形式要简单得多：

定理4·16 设定理4·15的条件都满足，则当$n_1 \to \infty$，$n_2 \to \infty$时，有

$$\sqrt{\frac{n_1 n_2}{n_1 + n_2}} S_{n_1 n_2}^{+} \xrightarrow{\mathscr{L}} K^{*}(x), \tag{4·200}$$

其中

$$K^{*}(x) = \begin{cases} 0, & \text{当 } x \leqslant 0; \\ 1 - e^{-2x^2}, & \text{当 } x > 0. \end{cases} \tag{4·201}$$

分布K^{*}的$100(1-\alpha)\%$分位点为$\left(\frac{1}{2}\log\frac{1}{\alpha}\right)^{1/2}$. 由此，当$n_1$和$n_2$都较大时，对给定的水平$\alpha$，可取以

$$\left\{ S_{n_1 n_2}^{+} > \sqrt{\frac{n_1 + n_2}{n_1 n_2}} \left(\frac{1}{2}\log\frac{1}{\alpha}\right)^{1/2} \right\}$$

为否定域的检验。当n_1, n_2较小时，否定域的临界值可根据$S_{n_1 n_2}^{+}$的分布定出来。

二、Колмогоров 检验：理论分布已知时。

设$X_1, \cdots, X_n$是从某总体中抽出的简单样本，要据以检验假设

$$H\text{：总体分布为 } F, \tag{4·202}$$

此处F为一已知分布，常称为理论分布。这是因为，分布F往往是根据某种理论、学说之下应有的分布，而$X_1, \cdots, X_n$则是实验的结果。检验假设（4·202），就从一个方面检验了该理论或学说

是否正确。在实用上，也常说实验数据 $X_1,\cdots,X_n$ 与理论分布 F 符合得怎样，故这类检验也称为拟合优度检验。

以 $F_n(x)$ 记 $X_1,\cdots,X_n$ 的经验分布函数，如（4·202）正确，则 F_n 作为 F 的估计，应与 F 相差不多。故引进统计量

$$S_n=\sup_{-\infty<x<\infty}|F_n(x)-F(x)| \tag{4·203}$$

Колмогоров 在 1933 年引进了基于 S_n 的检验，以

$$\{S_n>C\} \tag{4·204}$$

为否定域。与 Смирнов 统计量不同，Колмогоров 统计量 S_n 并非秩统计量。但不难证明：在理论分布 F 处处连续的假定下，当原假设（4·202）成立时，S_n 的分布与 F 无关，因此 C 之值只取决于样本大小 n 及给定的水平 α。这个事实的证明留给读者（习题21）。对较小的 n，C 之值可由 S_n 的精确分布定出。如 Miller的《Table of Percentage Points of Kolmogorov Statistics》(J. Amer. Statist. Assoc. 15, (1956), p. 111-121) 给出了 $n\leqslant100$ 时，$\sqrt{n}\,S_n$ 的90%，95%，99%分位点。也可参看中国科学院应用数学研究所概率统计教研室编的《常用数理统计表》。对较大的 n，可使用 Колмогоров 证明的下述极限定理：

定理4·17 设 F 在 $-\infty<x<\infty$ 处处连续且（4·202）成立，则当 $n\to\infty$ 时有

$$\sqrt{n}\,S_n\xrightarrow{\mathscr{L}}K(x) \tag{4·205}$$

其中 $K(x)$ 见（4·197）。

与 Смирнов 定理一样，这个定理也有好些记法，但没有一个容易的，在此只得从略。据这个定理，当水平取为 $\alpha=0.05$ 或0.01时，(4·204)中的临界值 C 可分别取为 $1.358/\sqrt{n}$ 或 $1.628/\sqrt{n}$。

Колмогоров 检验的效率如何？这个问题有一些学者研究过，其内容过于专门，在此我们只引述几条结论。一般的印象是：这个检验的性能是好的。

1. 通常用于检验（4·202）的检验，是 χ^2 拟合优度检验，一般说来，Колмогоров 检验与 χ^2 检验相比，在下述意义上处在有利地位：设G为一分布而记 $\Delta=\sup\limits_{-\infty<x<\infty}|F(x)-G(x)|$。$\Delta$ 可视为分布 F、G之间的距离，有时称为一致距离或 Колмогоров 距离。如果真正的理论分布为G，则假设（4·202）不成立。则我们希望否定原假设H，否定的概率愈大愈好。或反过来说，在具有一定的否定概率(即检验的功效）之下，Δ 愈小愈好。因 Δ 愈小，表明该检验能分辨出更小的（与H的）差异。在这方面，Колмогоров 检验优于 χ^2 检验，例如，在 $n=100$，$\alpha=0.05$ 而功效为 0.5 时，Колмогоров 检验能分辨的 Δ，可达到 χ^2 检验能分辨的 Δ 的二分之一（这只是一个大致的结论。因为无论是 Колмогоров 检验还是 χ^2 检验，其功效都不仅依赖于 Δ）。

Колмогоров 检验与 χ^2 检验相比还有其另外的优点，即 χ^2检验要把（$-\infty$，∞）分为若干个区间。区间数目及起迄点都有任意性。故同一组数据，由不同的人用 χ^2 检验去做，可以由于分组不同而得出不同之结果(即一个否定H，一个接受H)。Колмогоров 检验则没有这个随意性。另外，当 n 较小时，Колмогоров 检验的临界值C，是由在原假设下 S_n 的精确分布算出，有表可查而比较准确。χ^2 检验当 n 不大时，精确分布未知，而检验的临界值也系由其极限分布算出，故只是近似的。

2. 还可以拿 Колмогоров 检验与在特定情况下的最优检验相比，看其差距如何，以此得出其优良性的某些概念。举一例言之。设（4·202）中的理论分布F就是标准正态分布 $\Phi(x)$，而设想可能的对立假设是 $\Phi(x-\theta)$，$\theta>0$。这问题可用 Колмогоров 检验做，也可用通常的u检验做，其水平 α 的否定域为 $\sqrt{n}\,\bar{X}>u_\alpha$，$\bar{X}=\sum\limits_{i=1}^{n}X_i/n$。在假设检验理论中证明了：在所设情况下，这个$u$检验是水平 α 的一致最优检验。在 $\theta=1.5$ 时，Колмогоров 检验的功效为 0.895（取水平 $\alpha=0.05$，下同），而u检验

的功效为0.956，二者相差0.061。就是说，因用 Колмогоров 检验而损失的功效，也不过6％多一点。

三、Колмогоров 检验：理论分布带参数时.

在实用问题中，人们常希望用正态模型去分析试验数据。但有时并无充分把握肯定，试验数据确是来自正态分布，而需要通过适当的检验去判定。这问题与（4·202）的差别，在于在原假设中并未规定理论分布的确切形式，而只要求它是某一分布族的一员 这样，此处的检验问题可提为：根据从一总体中抽出的简单样本 $X_1,\cdots,X_n$ 去检验假设

H：总体分布为 $N(\mu,\sigma^2)$，对某个 $\mu\in(-\infty,\infty)$ 和 $\sigma^2>0$。 （4·206）

一般，我们有一个包含实参数向量 θ 的分布族 $\{F_\theta:\theta\in\Theta\}$ 。要根据简单样本 $X_1,\cdots,X_n$ 去检验假设：

H：总体分布是 F_θ，对某个 $\theta\in\Theta$ （4·207）

直观上看，前面二段中的方法容易推广到这里：先用样本$X_1,\cdots,X_n$ 对参数 θ 作一估计，以 $\hat{\theta}_n=\hat{\theta}_n(X_1,\cdots,X_n)$ 记估计量，如果样本确系抽自分布族$\{F_\theta:\theta\in\Theta\}$，则 $\hat{\theta}_n$ 接近 θ，因而 $X_1,\cdots,X_n$ 的经验分布 $F_n(x)$应接近于 $F_{\hat{\theta}_n}(x)$。故令

$$T_n=\sup_{-\infty<x<\infty}|F_n(x)-F_{\hat{\theta}_n}(x)|,$$

而当
$$T_n>C \qquad (4\cdot208)$$
时否定原假设(4·207)。例如当$\{F_\theta:\theta\in\Theta\}$ 为正态分布族时，可用样本均值 $\overline{X}$ 和样本方差 S^2 去估计分布族中的参数 μ 和 σ^2，然后计算

$$T_n=\sup_{-\infty<x<\infty}\left|F_n(x)-\Phi\left(\frac{x-\overline{X}}{S}\right)\right| \qquad (4\cdot209)$$

Φ 为 $N(0,1)$的分布．当 $T_n>C$时否定原假设。

这在形式上好像只是(二)段的简单推广，实际上问题复杂得多。问题就在统计量 T_n 的分布上。在理论分布完全已知时，(4·203）所定义的统计量 S_n，在F连续时，为“分布无关”的．此

处则不然，T_n 在原假设之下的分布，一依赖于 θ 的估计量 $\hat{\theta}_n$ 如何取。如在（4·209）中，你也可以用样本中位数 m 代 $\bar{X}$，所得统计量的分布不同。二依赖于理论分布族。如理论分布族为正态族，为负指数族，为 Cauchy 分布族时，产生的统计量 T_n 之分布不同。第三，T_n 还可能依赖于参数 θ 之具体值。如果情况确如此，则在给定水平 α 后，可能无法定出（4·208）中的常数 C，使检验具有水平 α。另外，T_n 的极限分布也可以依赖于此三者，而且很难求。例如，即使对最重要的正态分布族，由（4·209）定义的统计量 $\sqrt{n}T_n$ 的极限分布虽存在，但很不易求。

但容易证明：只要总体分布族确是正态族，则（4·209）所定义的 T_n 的分布，并不依赖于未知参数 μ 和 σ^2，因此，原则上可定出常数 C，使检验（4·208）有给定的水平 α。Lilliefors 在 1967 年用随机模拟法在 n 较小时，对 $\alpha=0.05$ 和 0.01 定出了临界值 C，如下表所示(表出给出的是 $100C$ 之值)：

n	5	6	7	8	9	10	11	12	13
0.05	33.7	31.9	30.0	28.5	27.1	25.8	24.9	24.2	23.4
0.01	40.5	36.4	34.8	33.1	31.1	29.4	28.4	27.5	26.8
n	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**25**	**30**
0.05	22.7	22.0	21.3	20.6	20.0	19.5	19.0	17.3	16.1
0.01	26.1	25.7	25.0	24.5	23.9	23.5	23.1	20.0	18.7

在 $n>30$ 时，（4·208）中的 C 可近似地取为 $0.866/\sqrt{n}$（相应于 $\alpha=0.05$）或 $1.031/\sqrt{n}$（相应于 $\alpha=0.01$）。Lilliefors 在其 1967 年的工作(见 J.Amer.Statist.Assoc. 62 (1967), p.399-402）中，还把这检验(用于正态场合)与 χ^2 拟合优度检验作了比较.得出的印象是：Колмогоров 检验优于 χ^2 检验。另外，Stephens 在 1974 年也提出了 C 的一个较好的近似值，为 $0.895/l_n$ ($\alpha=0.05$）和 $1.035/l_n$ ($\alpha=0.01$)，其中 $l_n=\sqrt{n}-0.01+0.85/\sqrt{n}$。

除正态族外，另一个重要的分布族是以

$$f_\theta(x)=\frac{1}{\theta}e^{-x/\theta}I(x>0),\ \theta>0 \qquad (4\cdot210)$$

为密度的负指数分布族。θ 作为这分布的期望，用样本均值 $\bar{X}$ 去估计。于是，为了检验"样本 $X_1,\cdots,X_n$ 来自分布族（4·210））"这个假设，计算

$$T_n=\sup_{x>0}|F_n(x)-(1-e^{x/\bar{X}})| \qquad (4\cdot211)$$

然后在 $T_n>C$ 时否定原假设。Lilliefors 在 1969 年也曾像对待正态族那样，通过随机模拟的办法去决定临界值 C。后来，到 1975—1976 年，Durbin 和 Margolin 等得到了(4·211)所定义的 T_n 在原假设成立之下的精确分布，因而可用来决定 C 之值。可参看 Durbin 在《Biometrika》62（1975）p.5—22 中所提供的表。当样本大小 n 充分大时，对 $\alpha=0.05$ 和 0.01，C 之值可近似地取为 $1.075/\sqrt{n}$ 及 $1.274/\sqrt{n}$。

习　　题

4-1　设 $X_1,\cdots,X_n$ 是从一维连续分布中抽出的简单样本，R_i 为 X_i 的秩，$i=1,\cdots,n$。试求 $E(R_i|X_1=x)$，x 给定。分 $i=1$ 和 $i\neq1$ 两种情况做。

4-2　设 $X_1,\cdots,X_m$ 和 $Y_1,\cdots,Y_n$ 分别是从一维连续分布 F 和 G 中抽出的简单样本，且合样本 $X_1,\cdots,X_m$，$Y_1,\cdots,Y_n$ 相互独立。以 R_1 记 X_1 在合样本中的秩。试求 R_1 的分布。本题说明：在样本非独立同分布时，秩的分布很复杂。

又：在 F 和 G 分别是（0,1）和（0,2）区间内的均匀分布时，计算数字结果。

4-3　试由定理4.3推出定理4·2。

4-4　验证以下两个例子，它们说明：若定理 4·4 的条件(1)和(2)那怕有一个不成立，则定理的结论可以不成立。

a.　$(c_{n1},\cdots,c_{nn})=((n-1),-1,-1,\cdots,-1)$，$\varphi(x)=x$（条件(1)不成立）。算出 $(L_n-l_n)/\sigma_n$ 的极限分布的形式。

b. $(c_{n1},\cdots,c_{nn})=(0,\cdots,0,1,\cdots,1)$，$\left[\frac{n}{2}\right]$个 0，其余为 1（$[a]$为不超过 a 的最大整数），$\varphi(x)=e^{1/x^2}$，$0<x<1$。（条件(2)不成立）

4-5　在 $\varphi(u)=u^2$，$0<u<1$ 这个特例，由定理 4·4 推出定理4·5。

4-6　设 $X_1,\cdots,X_n$ 为抽自分布F的简单样本，F在 0,1 两点不连续，其跳跃分别为$\frac{1}{3}$及$\frac{1}{4}$，F在其他点连续．以A记"不存在长大于1 的结"这个事件．计算其概率 $P(A)$。

4-7　记号同上题，但设F只有唯一的不连续点 0，其跳跃为$\frac{1}{3}$．记

$$\xi=\begin{cases}\text{结的长度，若存在长大于 1 的结；}\\ 0\quad\text{，其他情况，}\end{cases}$$

计算 ξ 的期望与方差．

4-8　写出 Mood 检验统计量(见 §4·2(一)段,2)在原假设下(且假定总体分布处处连续) 的渐近正态定理．又若把 Mood 统计量中的计分改为$\left(i-\frac{n}{2}\right)^2$，怎样在定理 4·4 的基础上，推出这个改变后的统计量的渐近正态性?

4-9　考虑例 4·6．证明：适当选择F，ARE $(W^+,\ t\ ;\ F)$可以取任意大的值。

4-10　利用公式（4·94)，计算 $\varphi(u)=u^m$ 时的检验的效率因子，及其与 Wilcoxon 检验的 ARE。

4-11　在 $n_1=3$，$n_2=2$，X_1，X_2，X_3，Y_1，Y_2 独立同分布，其公共分布函数为

$$F(x)=\begin{cases}0, & x\leqslant 0\\ x, & 0\leqslant x<1/2\\ x+1/3, & 1/2\leqslant x<2/3\\ 1, & x\geqslant 2/3\end{cases}$$

用平均法定结内变量的秩。以W记 Wilcoxon 秩和统计量($(Y_1, Y_2$ 之秩之和)。计算 $P(W=8)$(本例说明，当结存在时，秩统计量分布的计算很复杂)

4-12 在原假设成立时，且设总体分布F处处连续，计算由(4·109)式定义的多样本检验统计量 T_n 的期望。并据计算结果说明乘数因子$(n-1)/D_n$ 的理由。

4-13 证明(4·119)式

4-14 沿用(4·123)式中的记号。令

$$V_{jn}=\sum_{i=1}^{j-1}R_n(i,j),\quad j=2,\cdots,m$$

证明：对任何r，$3\leqslant r\leqslant m$，V_{rn} 与$(V_{2n},\cdots,V_{r-1,n})$ 独立。由此立即推出：$V_{2n},\cdots,V_{mn}$ 相互独立。利用这后一结论证明(4·125)。

4-15 在原假设成立之下，计算(4·131)定义的统计量 Q_n 的期望，以此说明乘数因子 $12n/(m(m+1))$的理由

4-16 设法把(4·139)的 T_n 表为 n 个 iid. 变量之和，因而证明(4·141)。

4-17 在§4·3，(二)4 中，设在模型

$$X_{ij}=\mu+\alpha_i+\beta_j+e_{ij},\quad i=1,\cdots,m,\ j=1,\cdots,n$$

中，所有的 e_{ij} 独立同分布，其公共分布连续，且原假设 $\alpha_1=\cdots=\alpha_m$ 成立。以 R_{ij} 记 $X'_{ij}=X_{ij}-X_j$ 在 $\{X'_{ij}: i=1,\cdots, m, j=1,\cdots,n\}$中的秩，则虽然$\{X'_{ij}\}$并非独立同分布，$\{R_{ij}: i=1,\cdots,m, j=1,\cdots,n\}$取(1，2，$\cdots mn$)的任一置换的概率仍为$1/(mn)!$。利用这个事实(先证明这个事实)推出(4·143)。

4-18 证明例 4·12 中的统计量 W^+ 在 $\theta=0$ 时的分布关于 $n(n+1)/4$ 点对称。

4-19 证明(4·192)式

4-20 证明位置参数的 HL 估计的平移同变性等两个性质(写在定理(4·14)之前)。

4-21 证明：在原假设成立且总体分布F处处连续时，

Колмогоров 统计量（4·203）的分布确与F无关．并算出当$n=2$时这分布的确切形式．

4-22　设总体分布F为离散分布，其在0，1两点处的概率都是$\frac{1}{2}$．问在原假设成立之下，Колмогоров 统计量的确切分布如何？又问在这一特殊情况下，Колмогоров 检验与χ^2拟合优度检验的关系如何？

4-23　证明：当原假设成立（即总体分布为正态$N(\mu,\sigma^2)$）时，$\sqrt{n}T_n$的极限分布，其中T_n由（4·209）定义，与μ和σ^2无关．

第五章　置换检验

§5·1　基本概念与例子

置换检验是条件检验这个更一般的概念的特例，但也是最重要和应用最广的特例，因此我们先得把“条件检验是什么”这个问题解释清楚.

让我们先看一个例子.

在§4·4的一段中，讨论过利用游程去检验随机性的方法，在样本 $X_1,\cdots,X_n$ 只取0、1这两个值时，以 ξ 记序列 $X_1X_2\cdots X_n$ 中，1游程的个数. 当 ξ 小于某个 C 时，就否定原假设. 为了找出 C，按照假设检验的一般步骤，应先确定在原假设下 ξ 的分布. 当原假设成立时，$X_1,\cdots,X_n$ 为独立同分布，其公共分布为

$$P_p(X_1=1)=p,\ P_p(X_1=0)=1-p,\ 0\leqslant p\leqslant 1. \quad (5\cdot 1)$$

此处 p 未知. 故原假设下不止一个分布，而是包含一个实参数 p 的一族分布（5·1）. 当 $X_1,\cdots,X_n$ 中恰有 m 个1时，事件 $\{\xi=k\}$ 的（条件）概率已在（4·153）中求得，现记为 $q(k|m)$. 在原假设下事件 $\{X_1,\cdots,X_n$ 中有 m 个1$\}$ 的概率为 $\binom{n}{m}p^m(1-p)^{n-m}$. 于是，由全概率公式，得到 ξ 在原假设下的分布为

$$P_p(\xi=k)=\sum_{m=0}^{n}\binom{n}{m}p^m(1-p)^{n-m}q(k|m),\ k=1,2,\cdots,n \quad (5\cdot 2)$$

此分布与 p 有关. 我们无法定出一个常数 C，使对一切 $p\in[0,1]$ 有 $P_p(\xi<C)=\alpha$，即使用随机化检验法也不行. 诚然，我们可以找到一个随机化检验，即“当 $\xi>1$ 时接受原假设，当 $\xi=1$ 时，以概率 $1-\alpha$ 接受原假设”，这检验正好有给定的水平 α（请读者自己验证），但这个检验显然没有什么用. 问题根本困难之点在于：

当ξ很小时，有两种可能：一是原假设确不成立（有某种相关性导致ξ很小），一是原假设其实成立，只是由于（5·1）中的p很接近0或1，使0，1中有一种符号数目很少，从而导致ξ很小，我们无法知道，上述两种可能性那一种是现实的，因而无法判定是否该否定原假设.

但在§4·4的一段中的讨论中，并未出现上述困难．原因在于：我们用一种条件化的手续，绕过了（5·2），即绕过了ξ在原假设下的"无条件"分布．具体做法是这样的：一经得到样本$X_1,\cdots,X_n$，我们先把其中1的个数η数出来．设$\eta=m_1$，记$n-m_1=m_2$。我们把$\{\eta=m_1\}$作为一个条件，而去求ξ在这个条件下的条件分布，这就是（4·153）。在此非常重要的一点是：尽管ξ的无条件分布（5·2）依赖参数p，这个条件分布则不依赖它．我们就这个条件分布去定否定域的临界值C，则C摆脱了对p的依赖，而只取决于水平α及η之值m_1。这样，该检验在任何条件$\{\eta=m_1\}$之下都有"条件水平"α，因此其（无条件）水平也是α．要注意的是：在此C已不是一个仅由α决定的常数，它还和η的取值有关：$C=C(\eta,\alpha)$。因此，C也是一个统计量．

这一变化不止是形式上的，而是有其实值内容：在"无条件检验"中，否定域临界值C要求是一常数，因此它无法分辨前述两种情况——即到底是由于正相关还是由于p太接近0或1而导致ξ很小．在此则不然，C的界限不固定，要看序列$X_1,\cdots,X_n$中1的个数而定，如1的个数接近$n/2$，则C定得大些．若1的个数太少或太多，则C定得小些．这样，在这一"条件化"的运作中，我们已把p值是否接近0、1的影响考虑进来了．

总结一句．在§4·4的一段中的检验，就是以统计量ξ（1游程个数）为基础，并在统计量η的条件化（η:$X_1,\cdots,X_n$中1的个数）之下的条件检验．条件检验的作用在于：在原假设为复合的情况下（如本例），它有助于克服因检验统计量在原假设下的分布不定（即依赖于原假设中究竟那一个分布出现）而带来的决定否

定域临界值C的困难.

如果单从形式上看，你可以说条件检验和无条件检验其实是一回事.事实上，引进一个统计量$T=(\xi,\eta)$和集合$A=\{(u,v):u=1,2,\cdots,n,v=0,1,\cdots,n,u<C(v,\alpha)\}$.上述条件检验可表为：

当$T\in A$时否定原假设，不然就接受. （5·3）

由于A是一个只依赖于α的集合，并无随机性，故检验（5·3）也就是通常的(无条件)检验. 充其量我们只能说，这检验形式略复杂些，不是常见的$\{T<C\}$或$\{T>C\}$这种形式而已.

所以，条件检验与无条件检验的差别不在形式，而在于引出检验的思想.在本例中,关键之点在于引进另一统计量η并以它为工具，对ξ施行条件化，这种思想可用于很多问题，而帮助克服如本例中用无条件检验而产生的那种困难. 后面我们有很多例子来解释这一点. 现在把条件检验的一般定义陈述如下：

设有样本X. X可以是简单样本$X_1,\cdots,X_n$，或由几个简单样本组成的合样本，也可以有更复杂的构成. 总体分布记为F. 故F可以是简单样本$X_1,\cdots,X_n$的公共分布，或合样本中两部分(或多部分)的分布（F，G）等，或有其他更复杂的构成. 设$\mathscr{F}$为一个分布族，原假设为$H:F\in\mathscr{F}$.给定水平α.

定义5·1 设$(T,M)=(T(X),M(X))$为一统计量，满足条件：若H成立，则在给定$M(X)=m$的条件下，$T(X)$的条件分布只依赖于m而不依赖于总体分布F. 找统计量$C(M,\alpha)$，使$P(T>C(M,\alpha)|M=m)=\alpha$，对$M$的任何可能值$m$(因为$T$在给定$M$时的条件分布只依赖$M$的给定值，这种$C$存在).则检验:

当$T(X)>C(M(X),\alpha)$时否定H，不然就接受 （5·4）

称为原假设H的一个条件检验，它有水平α.

在定义中为确定计,把条件否定域写成$T>C(M,\alpha)$的形状.它当然也可以有$T<C(M,\alpha)$或其他更复杂的形状.

从上面游程检验的例子中看到，在定义中涉及的两个统计量T和M，T是作为衡量与原假设的差距而引进，一般是基于直观

或某种理论或类似问题的启发．M则是作为施行“条件化”而特为引进的，它必须适合定义中的要求．M的引进没有一般的方法可言，但从问题的形式上常能有所启发，以下在一些例子中会见到

前面说过，置换检验是条件检验的特例．其特殊性，就在这个统计量M的形式上：

定义5·2 若在定义 5·1 中，M 是通过某种置换手续而产生的统计量，则检验（5·4）称为置换检验．

故置换检验并非唯一特定的检验，而是一类检验．其多样性就在这“某种”置换手续上．一个由n个元组成的序列，经置换可产生 $n!$个序列．这可称为“全面”置换，即不受任何约束的置换．在特定的问题中，出于需要，可对施行的置换加以一些约束，这时能产生的序列就没有$n!$这么多，见以下的例子．

例5·1(四格表) 考虑一个 2×2 列联表

B \ A	A_1	A_2	和
B_1	X_1	X_2	M_1
B_2	X_3	X_4	M_2
和	M_3	M_4	n

（5·5）

A,B 是一总体中的个体的两个属性，各有两个水平，A_1 和A_2，B_1和B_2．现随机观察了 n 个个体，发现(A_1,B_1)一类的有 X_1个，等等，要据以检验 “A,B 两属性独立” 这个原假设．

此问题在 A，B 两属性可取任意个水平的一般情况下，曾在 §4·4 的二段中按多样本问题的方式处理过．这种处理把“行和”或“列和”(其中之一)视为固定已知的，故在某种意义上说，不失为一个条件检验．可是那里的做法仍须乞援于大样本分布．此处用条件检验的方法，并进一步把所有的 M_i($i=1,\cdots,4$) 都视为随机的，而导出精确(小样本)检验，这个方法在历史上源于R.A.Fisher．

我们假定：如原假设不对，则 A，B 属性呈现正相关.这意思是说，当B取 B_1(足标小)时，A也更倾向于取 A_1(小足标).B取 B_2时A也倾向于取 A_2。这样一来，若以$P(A_i|B_j)$记当B取 B_j时，A 取 A_i 的条件概率，则 $P(A_1|B_1)-P(A_1|B_2)$可以作为衡量原假设是否成立的一个指标；当 A,B 独立时，$P(A_1|B_1)=P(A_1|B_2)=P(A_1)$，此指标为 0，当 A,B 有正相关时此指标大于 0。从表上数据看出，$P(A_1|B_1)-P(A_1|B_2)$可以用 $X_1/M_1-X_3/M_2$ 去估计之．故从形式上看，以 $X_1/M_1-X_3/M_2>C$ 为否定域之检验是一合适的检验．麻烦的是，即使在原假设“A,B独立”成立之下，此统计量之分布并不唯一确定，而取决于 A、B 的边缘分布

$$P(A_1)=1-P(A_2)=p,\ P(B_1)=1-P(B_2)=q \qquad (5\cdot6)$$

中的参数 p,q 但是

$$X_1/M_1-X_3/M_2=(nX_1-M_1M_3)/M_1M_2.$$

若给定了 M_1 和 M_3，则 M_2 也随之确定，因而上式只依赖于X_1，且随 X_1 之增减而增减．故可取定义 5·1 中之M为(M_1,M_3)，T为 X_1，问题在于验证：当给定 M_1 和 M_3 时，X_1 的条件分布不依赖于（5·6）中的 p 和 q。这不难验证：

$$P_{pq}(M_1=m_1,\ M_3=m_3)=\sum_{i=0}^{n}P_{pq}(X_1=i,\ X_2=m_1-i,\ X_3=m_3-i,\ X_4=n-m_1-m_2+i). \qquad (5\cdot7)$$

在原假设成立时，$(X_1,\ X_2,\ X_3,\ X_4)$构成多项分布：

$$P_{pq}(X_i=c_i,\ i=1,\cdots,4)=\frac{n!}{c_1!c_2!c_3!c_4!}(pq)^{c_1}(\bar{p}q)^{c_2}(p\bar{q})^{c_3}(\bar{p}\bar{q})^{c_4}$$

此处 $\bar{p}=1-p$，$\bar{q}=1-q$，$c_1,\cdots,c_4$ 为和等于 n 的非负整数，以此代入（5·7)，得

$$P_{pq}(M_1=m_1,\ M_3=m_3)=\sum_{i=0}^{n}\frac{(pq)^i(\bar{p}q)^{m_1-i}(p\bar{q})^{m_3-i}(\bar{p}\bar{q})^{n-m_1-m_3+i}n!}{i!(m_1-i)!(m_3-i)!(n-m_1-m_3+i)!}$$

$$=p^{m_3}\bar{p}^{n-m_3}q^{m_1}\bar{q}^{n-m}\binom{n}{m_1}\sum_{i=0}^{n}\binom{m_1}{i}\binom{n-m_1}{m_3-i}$$

$$=p^{m_3}\bar{p}^{n-m_3}q^{m_1}\bar{q}^{n-m_1}\binom{n}{m_1}\binom{n}{m_3}$$

又

$$\begin{aligned}&P_{pq}(X_1=k,M_1=m_1,M_3=m_3)\\&\quad=P_{pq}(X_1=k,X_2=m_1-k,\ X_3=m_3-k,X_4\\&\quad=n-m_1-m_3+k)\\&\quad=\frac{n!}{k!(m_1-k)!(m_3-k)!(n-m_1-m_3+k)!}\\&\qquad(pq)^k(\bar{p}q)^{m_1-k}(p\bar{q})^{m_3-k}(\bar{p}\bar{q})^{n-m_1-m_3+k},\end{aligned}$$

由以上两式得

$$\begin{aligned}&P_{pq}(X_1=k\mid M_1=m_1,M_3=m_3)\\&=P_{pq}(X_1=k,M_1=m_1,M_3=m_3)/P_{pq}(M_1=m_1,M_3=m_3)\\&=\binom{m_1}{k}\binom{n-m_1}{m_3-k}\Big/\binom{n}{m_3},\end{aligned}$$

此式与 p,q 无关（注意这是在原假设成立的前提下．这一点不要忘记），因而符合定义 5·1 的条件．找 $C=C(m_1,m_3,\alpha)$，使

$$\sum_{k=c+1}^{n}\binom{m_2}{k}\binom{n-m_1}{m_3-k}\Big/\binom{n}{m_3}=\alpha, \qquad (5\cdot 8)$$

然后在 $X_1>C(M_1,M_3,\alpha)$ 时否定原假设．这个条件检验在原假设成立时确切地有水平 α，不管（5·6）中的 p,q 取怎样的值．如果不存在 C 使(5·8)成立，则须修改 α 之值，或使用随机化手续．

例5·2（两样本问题，成组比较）　设有治疗同一种疾病的两种药物 A 和 B．为比较其优劣，收集了 $n=n_1+n_2$ 个患者，随机地从中挑选 n_1 个服药 A，其余 n_2 个服药 B．假设治疗效果通过某项指标体现．n_1 个服药 A 后一段时期，量出其指标为 $X_1,\cdots,X_{n_1}$．服药 B 者指标为 $Y_1,\cdots,Y_{n_2}$．

问题的确切统计模型，随所作的假定而异．以下我们将分别考虑三种可能的提法．

1．按我们在以往几章中常用的一种提法，设 $X_1,\cdots,X_{n_1}$ 是

来自一个具分布 $F(x)$的总体，而 $Y_1,\cdots,Y_{n_2}$是来自一个具分布 $F(x-\theta)$的总体．F未知，θ为未知实参数．"药物A、B的疗效相同"的原假设，归结为 $\theta=0$，对立假设为 $\theta\neq0$。

此问题在上一章中已用秩方法处理过，此处我们用条件检验的方法去处理．据定义 5·1，在用条件法处理一个检验问题时，要引进两个统计量T和M。前者是作为衡量与原假设差距之指标，而后者是作条件化之用．就目前问题而言，T可选择为 $\bar{Y}-\bar{X}$。当$|T|>C$时否定原假设，可是即使在 $\theta=0$ 时，T之分布仍依赖F，因而无法找到常数C，使对一切F有 $P_F(|T|>C|\theta=0)=\alpha$。

为说明使用条件化即选择M之方法，要引进一些记号．首先，令$(Z_1,\cdots,Z_{n_1})=(X_1,\cdots,X_{n_1})$，$(Z_{n_1+1},\cdots,Z_n)=(Y_1,\cdots,Y_{n_2})$。$Z=(Z_1,\cdots,Z_n)=(X_1,\cdots,X_{n_1},Y_1,\cdots,Y_{n_2})$。对 n 维欧氏空间中的一点 $u=(u_1,\cdots,u_n)$，定义 $T(u)=\sum\limits_{i=n_1+1}^{n}u_i/n_2-\sum\limits_{i=1}^{n_1}u_i/n_1$。以$M(Z)$记集合①

$M(Z)=\{(Z_{i_1},\cdots,Z_{i_n}):(i_1,\cdots,i_n)$跑遍$(1,\cdots,n)$的一切置换$\}$，把$M(Z)$中的$n!$个点排列为 $Z^{(1)},\cdots,Z^{(N)}$($N=n!$．注意Z本身是这 $n!$ 个点之一)．当原假设 $\theta=0$ 成立时，$Z_1,\cdots,Z_n$为独立同分布，故在给定 $M(Z)$的条件下，$Z^{(1)},\cdots,Z^{(N)}$中每一个有同等的概率出现．由此可知，在给定$M(Z)$时，$T(Z)$的条件分布是

$$P(T(Z)=T(Z^{(i)})|M(Z))=1/N,\quad i=1,\cdots,N \tag{5·9}$$

(如$T(Z^{(i)}),\cdots,T(Z^{(N)})$中有相同的，则概率要合并)这个条件分布与$F$无关，因而适合定义5·1的要求．

现在可以根据(5·9)，把使用(T,M)对原假设 $\theta=0$ 进行条件检验的步骤列举如下：

1．由Z出发，经一切可能的置换，得 $Z^{(1)}\cdots,Z^{(N)}$。

---

① 也可以把$M(Z)$定义为$Z_1,\cdots,Z_n$的次序统计量，这形式较简单，但此处的定义对Z_i为高维时也适用。

2. 计算N个值$|T(Z^{(i)})|$, $i=1,\cdots,N$，把它们按由大到小排列为 $t_1\geqslant t_2\geqslant\cdots\geqslant t_N$。

3. 计算 $N_\alpha=N\alpha$，α为给定的水平. 计算$T(Z)$. 若$|T(Z)|\geqslant t_{N\alpha}$，则否定原假设. 不然就接受原假设. 当$N\alpha$非整数时，要适当修正$\alpha$。

在本例及类似的问题中以上步骤可简化一些，因为$T(Z^{(i)})$之值，其实只依赖于 $Z^{(i)}$ 的最后n_2个分量究竟包含了Z中的那些分量. 故如$n_1=4$，$n_2=3$，则(1,7,2,4,3,5,6)这个置换与(7,4,1,2,5,6,3)这个置换所导致的T值一样. 因此，最多只有$\binom{n}{n_2}$个不同的T值，它取决于置换$(i_1,\cdots,i_n)$中后n_2个元构成的子集. 为明确概念，举一个数字例子. 设$n_1=3$，$n_2=2$，$Z=(1\cdot5,1\cdot3,2\cdot1,2\cdot4,2\cdot7)$以$(j_1,j_2)$记最后两位为$j_1$和$j_2$的那种置换，则对一切可能的置换.$T(Z^{(i)})$至多只取如下10个相异值：

(1, 2)：T值为$\dfrac{1\cdot5+1\cdot3}{2}-\dfrac{2\cdot1+2\cdot4+2\cdot7}{3}=-1\cdot000$

(1, 3)：T值为$\dfrac{1\cdot5+2\cdot1}{2}-\dfrac{1\cdot3+2\cdot4+2\cdot7}{3}=0\cdot333$

(1, 4)：T值为$\dfrac{1\cdot5+2\cdot4}{2}-\dfrac{1\cdot3+2\cdot1+2\cdot7}{3}=-0\cdot083$

(1, 5)：T值为$\dfrac{1\cdot5+2\cdot7}{2}-\dfrac{1\cdot3+2\cdot1+2\cdot4}{3}=0\cdot167$

(2, 3)：T值为$\dfrac{1\cdot3+2\cdot1}{2}-\dfrac{1\cdot5+2\cdot4+2\cdot7}{3}=-0\cdot500$

(2, 4)：T值为$\dfrac{1\cdot3+2\cdot4}{2}-\dfrac{1\cdot5+2\cdot1+2\cdot7}{3}=-0\cdot250$

(2, 5)：T值为$\dfrac{1\cdot3+2\cdot7}{2}-\dfrac{1\cdot5+2\cdot1+2\cdot4}{3}=0\cdot000$

(3, 4)：T值为$\dfrac{2\cdot1+2\cdot4}{2}-\dfrac{1\cdot5+1\cdot3+2\cdot7}{3}=0\cdot417$

(3, 5)：T值为$\dfrac{2\cdot1+2\cdot7}{2}-\dfrac{1\cdot5+1\cdot3+2\cdot4}{3}=0\cdot667$

(4，5)：T值为$\frac{2\cdot4+2\cdot7}{2}-\frac{1\cdot5+1\cdot3+2\cdot1}{3}=0.917$

就是说，当得到样本$Z=(1\cdot5,1\cdot3,2\cdot1,2\cdot4,2\cdot7)$时，在给定$M(Z)$的条件下，$T(Z)$的条件分布是以概率0.1取上述10个值的每一个．如若给定水平$\alpha=0\cdot3$，则由此条件分布，当$|T(Z)|\geqslant0\cdot667$时否定原假设．现有$T(Z)=0.917$，$|T(Z)|=0.917>0\cdot667$，故应否定原假设．

这样，我们在不对总体分布F作任何假定的情况下，作出了原假设的一个检验，其检验统计量在原假设成立时，对原假设上的分布为分布无关的．在某种意义上，这一检验比秩检验更为一般，因为此处无须假定总体分布F处处连续——在秩检验的场合，当允许总体分布不连续因而"结"出现时，需要把结统计量作为$M(Z)$来实行条件化，才能达到在原假设下分布无关的结果．

注意在本例中，作为条件化之用的统计量$M(Z)$是由样本Z经过一切置换而产生之集．凡是这样构造的条件检验就称为置换检验．这在定义5·2中已说明了．

2．现在考虑本问题的另一种模型．刚才讨论过的模型的背景是：参与试验的n_1+n_2个患者是从极大一批情况基本相似的患者中随机抽取的．这个极大的"患者总体"是产生分布F的依据．

现设由于条件的限制，我们只能就手头可获得的n_1+n_2个患者做试验．它们是我们所有的全部"试验材料"，其来历因而不能视为是从一大总体中随机抽得的．这样，前面讨论过的模型就不适用．

把这$n=n_1+n_2$个患者编号．作为一个假定，设A、B两种药在每一患者身上的效果之差是恒定的，这样，若患者i用药A后指标为a_i，则如他不用A而用B，指标当为$a_i+\theta$．现从这$n=n_1+n_2$个患者中随机地抽取n_2个让他们服药B，剩下n_1个服药A．其指标值仍记为$X_1,\cdots,X_{n_1}(A)$和$Y_1,\cdots,Y_{n_2}(B)$，$Z=$

$(Z_1,\cdots,Z_n)=(X_1,\cdots,X_{n_1},Y_1,\cdots,Y_{n_2})$.

在此，即使原假设 $\theta=0$ 成立，$Z_1,\cdots,Z_n$ 也不是独立同分布。但其分布为：Z 取$(a_1,\cdots,a_n)$ 的任一置换的概率都是 $1/n!$ 。但在 $\theta=0$ 时，由 $Z_1,\cdots,Z_n$ n个数构成的集合，与由 $a_1,\cdots,a_n$ 这 n 个数构成的集合一样．故在原假设 $\theta=0$ 成立时，任一统计量 $T(Z)$ 在给定 $M(Z)$ ($M(Z)$ 的定义同前）的条件下，其条件分布①仍有（5·9）式所标示的性质（$N=n!$），故前面用 $T(Z)=(Z_{n_1+1}+\cdots+Z_n)/n_2-(Z_1+\cdots+Z_{n_1})/n_1$ 而作的检验法，一字不改地移到此处．

这两种模型那一种更合理？这当然要看样本（n 个患者）是如何得来的．不过可以注意：即使样本确是从一大总体中随机抽来，用第二种模型去处理也不错(反过来则不行)，这是因为，在第二种模型中，对样本的来源毫无条件，故即使是随机抽来的也无所谓．而且，从事实的角度看，往往是第二种模型更符合情理．因为在做这类试验中，往往只能"就地取材"，而未必有机会在一个很大的范围内去随机挑选．

顺便说一句：这第二种模型更好地体现了 Fisher 的试验设计三原则之一——随机化原则．事实上，正是随机化原则的使用(即从 n 个患者中"随机地"挑选 n_2 个服药B这个手续)，赋予统计量 $T(Z)$一定的概率分布，因而可能用统计的方法去处理之．如果不用随机化，则我们无法判断：Y平均与X平均之间的差异，究竟是因A，B的差异而来，还是由于个体之间的差异而来．至于第一种模型，其随机结构（体现在分布 $F(x)$及 $F(x-\theta)$ 中）已由样本是从一个极大的总体中抽取这个背景而确定了．在这个背景之下，看不出从已有的 n 个患者中再用随机化方法去抽 n_2 个一举有何作用，因为它未在原有的随机化结构上添加任何新东

①读者一定注意到，在这第二种模型下，当原假设成立时，$M(Z)$ 并无随机性．故$T(Z)$在给定$M(Z)$之下的条件分布，即等于$T(Z)$之无条件分布，所以，在这模型下作的置换检验，并无条件检验的气味．

西.

这后面一点也就是很多古典方差分析模型之难于自圆其说的所在．拿常见的随机区组设计来说，m 个品种在 v 个区组（每区组包含 m 小区）中施行随机化．模型是

$$X_{ij}=\mu+a_i+b_j+e_{ij},\ i=1,\cdots,m,\ j=1,\ \cdots,v, \quad (5\cdot10)$$

a_i,b_j 分别为品种 i 和区组 j 的效应．至于区组内各小区，则在模型中假定为绝对均匀的，因此在（5·10）中，并无体现“小区效应”之项．可是既然同一区组内各小区绝对均匀，那么在区组内施行随机化还有何必要？对此，一般作的解释是：“尽管同一区组内各小区已很均匀，但还是有些差异的．为防止系统偏差，故在区组内施行随机化云云”．此说显得有点矛盾：既承认小区之间不均匀，为何不体现在模型中？以后我们将看到：在类似于上面第二个模型那种解释下，把小区之间的差异考虑进来，而且，（5·10）中的误差项 e_{ij}，正好就来源于这种差异，这种解释就显得很自然而合理。

3．本问题另一种可以考虑的模型为：若在第 i 个患者身上施药 A，则其指标为 a_i+e_i，若施药 B 则为 $a_i+\theta+e_i$， $i=1\cdots,n$。这里，a_i 或 $a_i+\theta$ 这一项的含义，与 2 中的模型相同．而 $e_1,\cdots,e_n$ 为独立同分布的随机变量，它反映了药的疗效中与患者个体无关的那一部分，如剂量大小有随机性波动．环境因素以至测量误差等．从事理上分析应该说，这是与实际情况最接近的一种模型．从对称性考虑读者容易理解：若仍按前述设计，从已有的 n 位患者中随机挑选 n_2 位施药 B，而保持前面的一切记号，则在给定 $M(Z)$ 的条件下，$T(Z)$ 的条件分布仍如（5·9）所示．所以，虽然模型变了，前面描述过的条件检验步骤，可以一字不改地移于此处．

从本例可以看出施行置换检验的一个实际困难所在，即计算量很大．如本例中若取 30 位患者，各一半服药 A 和 B．则需要计算

$\binom{30}{15}=5348880$

个不同的 T 值再排序．若 $n_1=n_2=50$，则将成为天文数字．为克服这种困难又须乞援于极限分布，下两节将处理这个问题．

例5·3(成对比较试验) 为估量两个种子品种 A,B 是否在产量上有显著差异，选择 $2n$ 块大小形状一样的地块并将其结成 n 组，每组两块．在分组时，使每组内的两块地在条件上尽可能接近．不同组内的地块条件可以有较大差异．

在这 n 组内各自独立地施行随机化，从两块中抽一块用品种 A，剩下那块用品种 B．把第 i 组内用品种 A 那块地的亩产记为 X_i，用品种 B 的则记为 Y_i，$i=1,\cdots,n$．记 $T=\bar{Y}-\bar{X}$ 则 $|T|$ 可以作为衡量两品种是否有差异的指标：当 $|T|$ 大时，否定"两品种产量无差异"的原假设．至于这界限如何定，则要看取怎样的统计模型．在初等教本中，把同组内两块地看成绝对均匀，因而在 i 组中两块地亩产之差 Y_i-X_i 由两部分构成：一部分是 θ，反映品种 B 与 A 亩产差的理论值．另一部分 e_i 是随机误差．于是有

$$Y_i-X_i=\theta+e_i,\ i=1,\cdots,n$$

原假设"品种无差异"转化为 $\theta=0$．又进一步假定 $e_1,\cdots,e_n$ 独立同分布并有正态分布 $N(0,\sigma^2)$，则本问题可以用熟知的"一样本 t 检验"去处理．

与上例相似，这个模型存在一些问题．一是若不假定 e_i 有正态分布该怎么办．这可以用秩方法，例如符号检验或 Wilcoxon 符号秩和检验去处理．另一个带根本性的问题是：往往同一组内两个地块仍有些差异，不可忽略不计，假定为绝对均匀不合理．而且，既然已假定为绝对均匀，在同一组内施行随机化还有何必要，这是本模型无法自圆其说之处．

因此，我们采取类似于前例的模型 2 的做法．每一组内两块地各赋予一值 a_{i1} 和 a_{i2}．其意义是：若用品种 A，则这两块地的亩产分别为 a_{i1} 和 a_{i2}．若用品种 B 则分别为 $a_{i1}+\theta$ 和 $a_{i2}+\theta$．a_{i_1},a_{i_2}

都未知，它们反映地块本身的条件，在分组时我们要使两块地的条件尽可能均匀，因而a_{i1}和a_{i2}的差距尽量小，但以下的方法并不依赖这一点。不过，若分组不当而致使 a_{i1} 和 a_{i2} 有较大差距，将影响本方法的功效。

记 $Z_i=Y_i-X_i$，$i=1,\cdots,n$。设想 $\theta=0$（原假设成立），则 Z_i 只能取 $a_{i2}-a_{i1}$ 和 $-(a_{i2}-a_{i1})$两值，究竟取那一个，则要看在第 i 组内施行随机化的结果。由于各组独立地施行随机化，故$Z_1,\cdots,Z_n$ 独立同分布，且 Z_i 以$\frac{1}{2}$的概率取 b_i 及 $-b_i$，其中$b_i=a_{i2}-a_{i1}$。由此可知，在给定集合

$$M(Z)=\{(\pm Z_1,\pm Z_2,\cdots,\pm Z_n):\pm\text{号取一切可能}\} \quad (5\cdot11)$$

的条件下，这集合的 2^n个点中每一个点有同等的机会(概率 $1/2^n$)出现。把 $M(Z)$中的 $N=2^n$ 个点记为 $Z^{(1)},\cdots,Z^{(N)}$，$T(Z)=\bar{Z}$，则据上述有

$$P\{T(Z)=T(Z^{(i)})|M(Z)\}=1/N,\quad i=1,\cdots,N. \quad (5\cdot12)$$

因此分布与 a_{i1},a_{i2} 这些量无关，(T,M) 这一对统计量适合使用条件检验的需要，现可以把施行的具体步骤列举如下：

1. 得到 X_i,Y_i，$i=1,\cdots,n$ 后，算出 $Z_i=Y_i-X_i$，$i=1,\cdots,n$。

2. 列出 2^n 个点$(\pm Z_1,\pm Z_2,\cdots,\pm Z_n)$，将它排列为$Z^{(1)},\cdots,Z^{(N)}$，$N=2^n$。

3. 对每个 $Z^{(i)}$算出 $T(Z^{(i)})=Z^{(i)}$ 的 n 个分量的算术平均，$i=1,\cdots,N$。把这N个值的绝对值按由大到小排列为 $t_1\geqslant\cdots\geqslant t_N$。

4. 给定检验水平 α，算出 $N_\alpha=N\alpha$。算出$|T(Z)|$（按开头之记号为 $(\bar{Y}-\bar{X}|)$。若 $|T(Z)|\geqslant t_{N\alpha}$，则否定原假设。不然就接受原假设，如果 $N\alpha$ 不为整数，则需调整 α 之值，或使用随机化检验法。

为明确计看一个数字例子：设 $n=3$，在三个组内试验所得 X和Y 值分别为

$X_1=5.1$, $Y_1=5.5$; $X_2=5.4$, $Y_2=4.9$; $X_3=5.2$, $Y_3=5.1$
由这些值算出 $(Z_1,Z_2,Z_3)=(0.4, -0.5, -0.1)$. 由此出发产生的 $M(Z)$ 包含 8 个点:

$Z^{(1)}=(0.4, -0.5, -0.1)$, $Z^{(2)}=(0.4, -0.5, 0.1)$,

$Z^{(3)}=(0.4, 0.5, -0.1)$, $Z^{(4)}=(-0.4, -0.5, -0.1)$,

$Z^{(5)}=(0.4, 0.5, 0.1)$, $Z^{(6)}=(-0.4, -0.5, 0.1)$,

$Z^{(7)}=(-0.4, 0.5, -0.1)$, $Z^{(8)}=(-0.4, 0.5, 0.1)$,

由这 8 个点所标出的 $T(Z^{(i)})$值依次为

-0.067, 0.000, 0.267, -0.333, 0.333, -0.267, 0.000, 0.067,

其按绝对值大小依次排列之结果为

0.333, 0.333, 0.267, 0.267, 0.067, 0.067, 0.000, 0.000.

若取 $\alpha=1/4$,则 $N_\alpha=N\alpha=2$.只有在$|\bar{Z}|\geqslant 0.333$ 时才能否定"两品种无差异"的原假设. 现有 $\bar{Z}=-0.067$, $|\bar{Z}|<0.333$,不能否定原假设.

在本例中, $M(Z)$ 也是由置换产生的. 不过这个置换受到限制: 它不是在原始数据 $(X_1,Y_1,\cdots,X_n,Y_n)$ 中任意置换, 而只能在一对内作置换, 即 X_i,Y_i之间可交换位置, 但 X_i 和 X_j, 或 X_i 与 Y_j, $j\neq i$, 都不能交换. 从这两例也看出, 当在试验中施行随机化时, $M(Z)$ 如何产生直接由随机化的内容所决定.

在上例中 1、2 两个模型的比较上所说的话, 也完全适用于此处, 在此模型下, 同一组内两个小块间差异的作用有充分的体现. 本例实际上就是区组大小为 2 的完全随机区组设计. 以后要考虑一般情形.

例5·4(多样本问题, 一元方差分析设计) 这就是把例 5·2 中两种药物 A、B 的比较问题, 推广为$A_1,\cdots,A_c$等 c 种药物的比较问题. 设有 $n=n_1+\cdots+n_c$ 个患者参与试验. 将他们随机地

分为 c 组，分别包含 n_1、n_2、…，n_c 个人，使第一组的人服药 A_1，…，第 c 组的人服药 A_c。第 i 组 n_i 个人的指标记为 $X_{i1},\cdots,X_{in_i}$，$i=1,\cdots,c$。令 $Z=(Z_1,\cdots,Z_n)=(X_{11},\cdots,X_{1n_1},\cdots,X_{c1},\cdots,X_{cn_c})$。

与例5·2一样，有三种模型可选用·第一种是假定样本 $X_{i1},\cdots,X_{in_i}$ 来自分布 $F(x-\theta_i)$，$i=1,\cdots,c$。原假设"各药物之间效应无差别"归结为 $\theta_1=\cdots=\theta_c$。若进一步假定 $F(x)$ 为正态分布 $N(0,\sigma^2)$，则是初等教本中一元方差分析的典型提法，用熟知的 F 检验去处理之．若对 $F(x)$ 的形式不作假定，则此法不行，可用第四章讲述的秩方法处理，也可以用下面讲到的第二种模型去处理之。

第二种模型是假定每一患者有一个反映其条件的常数与之对应．这样，若在第 i 名患者身上施药 A_j，则其指标为 $a_i+\theta_j$，$a_1,\cdots,a_n$ 未知且可有差异．为施行条件化检验，需要两个统计量 T 和 M。M 的取法如前，即由 Z 的 n 个坐标置换而产生的由 $n!$ 个点构成的集，或简单地即 $Z_1,\cdots,Z_n$ 的次序统计量．至于 T，则须反映各药物间的差距，记

$$\bar{Z}_i=\sum_{j=n_1+\cdots+n_{i-1}+1}^{n_1+\cdots+n_i} Z_j/n_i,\quad i=1,\cdots,c,\ \bar{Z}=\sum_{i=1}^{n} Z_i/n. \qquad (5\cdot13)$$

$\bar{Z}_1,\cdots,\bar{Z}_c$ 分别反映药物 $A_1,\cdots,A_c$ 的平均效果．当原假设成立时，它们的值应比较接近，反之则有较大差距。故可以用加权和

$$T=T(Z)=\sum_{i=1}^{c} n_i(\bar{Z}_i-\bar{Z})^2 \qquad (5\cdot14)$$

作为衡量试验数据与原假设差异的指标。

以下的步骤即与例 5·2 无异：得出 Z 后，经置换得出 $N=n!$ 个点 $Z^{(1)},\cdots,Z^{(N)}$。算出 N 个值 $T(Z^{(i)})$，$i=1,\cdots,N$，把它们按由大到小排列为 $t_1\geqslant\cdots\geqslant t_N$。算出 $T(Z)$。给定水平 α，当 $T(Z)\geqslant t_{N\alpha}$ 时否定原假设。当然，你也可以用别的统计量（直观上看来合理者）代替 T。检验步骤无异．与例 5·2 相似，这里并

无必要计算 $n!$ 个不同的 T 值. 其实相异者至多不超过 $n!/(n_1!n_2!\cdots n_c!)$ 个.

第三种模型在提法和处理方法上也与例 5·2 相似，此处不重复了.

例5·5(独立性检验) 设(X_i,Y_i)，$i=1,\cdots,n$，是二元总体(X,Y)的简单样本，要检验"X,Y独立"这个假设.

此问题在第四章中曾用秩方法处理过，若假定(X,Y)的联合分布为正态，则可以通过相关系数

$$r=\left(\sum_{i=1}^{n}X_iY_i-n\bar{X}\bar{Y}\right)\Big/\left(\sum_{i=1}^{n}(X_i-\bar{X})^2\sum_{i=1}^{n}(Y_i-\bar{Y})^2\right)^{1/2}$$

去检验之，但当(X,Y)之分布不假定为正态时，r的(无条件)分布定不出来，而此法不通，可用条件检验方法去处理之. 记$M=(M_1,M_2)$,其中 M_1 和 M_2 分别是 $X_1,\cdots,X_n$ 的次序统计量和$Y_1,\cdots,Y_n$的次序统计量. T就选择为 r.

在给定 M_1 和 M_2后，$(X_1,\cdots,X_n)$和$(Y_1,\cdots,Y_n)$分别各有 $n!$ 种置换，但 r 的分母与这置换无关，分子中的 $n\bar{X}\bar{Y}$ 也与这置换无关，它们在给定(M_1,M_2)的条件下为常数. 只有$\sum_{i=1}^{n}X_iY_i$一项可随这置换而变化，但也只能产生 $n!$ 个不同之值——因为若 $X_1,\cdots X_n$和$Y_1,\cdots,Y_n$经受同一置换,则不改变$\sum_{i=1}^{n}X_iY_i$之值，故不妨设 $X_1,\cdots,X_n$ 固定这次序不动,而只有$Y_1,\cdots,Y_n$ 作置换，这样产生 $n!$ 个值，因为在原假设成立时 $X_1,\cdots,X_n$ 和 $Y_1,\cdots,Y_n$ 都是独立同分布. 在给定 $M=(M_1,M_2)$ 的条件下,这 $n!$ 个值有等概率 $1/n!$. 由以上考虑，得出检验的步骤如下：

1. 就 $(1,2,\cdots,n)$的每一个置换$(i_1,i_2,\cdots,i_n)$计算$\sum_{j=1}^{n}X_jY_{i_j}$之值，把这 $n!$ 个值记为 $t_1,\cdots,t_N$，$N=n!$；

2. 算出 N 个值 $t_i'=|t_i-n\bar{X}\bar{Y}|$，$i=1,\cdots,N$，并把它们按由大到小排列为 $t_{(1)}'\geqslant t_{(2)}'\geqslant\cdots\geqslant t_{(N)}'$；

3. 给定水平 α. 算出 $\left|\sum_{i=1}^{n}X_iY_i-n\bar{X}\bar{Y}\right|$（事实上它是上述 N 个值中之一）。若此值 $\geqslant t_{(N\alpha)}$，则否定原假设，不然就接受原假设。

这个条件检验在不对（X,Y）的分布作任何假定的情况下，给出确切的水平 α。这是本节的几个例子所体现出的一个共同优点．置换检验之所以没有得到广泛应用，其原因也可以在这几个例子中看到，即计算量太大．为克服这一困难，当样本大小较大时，有必要乞援于大样本理论，而这又会导致回到传统检验法．

例 5·6（概率变点问题） 我们再来给出条件检验的一个有趣的应用．设定时地观察某事件 A 是否发生．设开始时，A 的概率稳定在 p_1。到某个未知的时刻，它可以突变到另一个值 p_2。当 $p_1\neq p_2$ 时这个时刻就称为（概率）变点．考虑至多只含一个变点的情况，统计模型可表为：有独立样本 $X_1,\cdots,X_n$，分布是

$$P(X_i=1)=1-P(X_i=0)=p_1,\ i=1,\cdots,m-1$$
$$P(X_i=1)=1-P(X_i=0)=p_2,\ i=m,\cdots,n.$$
$$(p_1,p_2,m\text{ 未知})$$

要依据样本检验"变点不存在"即 $p_1=p_2$ 这个原假设 H，以及当 H 被否定时，估计变点 m。

记 $U_k=X_1+\cdots+X_k$，$V_k=k-U_k$. U_k 是到时刻 k 为止 A 的累计出现次数，故下文的方法称为累计次数法。定义统计量

$$T_k=k(U_k/k-U_n/n),k=1,\cdots,n,$$

易见

$$E(T_k)=\begin{cases}kn^{-1}(n-m+1)(p_1-p_2),\text{当}1\leqslant k\leqslant m-1;\\(n-k)n^{-1}(m-1)(p_1-p_2),\ \text{当}m\leqslant k\leqslant n,\end{cases}$$

由此式看出：只要 $p_1\neq p_2$，则 $|E(T_k)|$ 开始随 k 增加而增加，到 $k=m-1$ 处达到最大，然后随 k 增加而下降．而当 $p_1=p_2$ 时则总为 0. 这个事实启发了以下的检验法：令 $T=\max(|T_1|,\cdots|T_n|)$。当 T 大于某常数 C 时，否定原假设 H. 不然就接受 H。由于 T 在

H 成立时之分布既复杂且与 p_1, p_2 之公共值 p 有关，C 不易定出。但我们可证明：在给定 $U_n = n_1$ 的条件下，$n_1^{-1}n_2^{-1}nT$ 的条件分布，与在原假设(两分布同)成立之下，样本大小分别为 n_1 及 $n-n_1=n_2$ 的 Смирнов统计量的分布相同。由于在 n_1, n_2 较小时 Смирнов 统计量的分布有表可查且当 n_1, n_2 大时定出了其极限分布，故可凭借这个关系来检验 H。

为证明这一事实，只须注意，若将 T_k 改写为 $T_k = n_1(U_k/n_1 - k/n) = n_1(U_k/n_1 - U_k/n - V_k/n)$，即得 $|T_k| = n^{-1}n_1n_2|U_k/n_1 - V_k/n_2|$。因而 $nn_1^{-1}n_2^{-1}T = \max\limits_k |U_k/n_1 - V_k/n_2|$，其结构正好与 Смирнов 统计量同。只在后者而言，X 样本所占的 n_1 个位置，现在由 n_1 个 1 占据。当 Смирнов 原假设成立时，n_1 个 X 样本在全部的 n 个样本中所占位次，在 $\binom{n}{n_1}$ 种方法中为等可能。而此处在 $p_1 = p_2$ 时，在给定 $U_n = n_1$ 的条件下，n_1 个 1 所占位次，在全部 $\binom{n}{n_1}$ 种方法中也是等可能。不难理解，这就证明了所要的结果。

§5·2 大样本置换检验

置换检验，或一般地说条件检验，是基于统计量 T 在给定统计量 M 的条件下的条件分布。这个分布往往是一个包含大量数值的离散型分布，计算和应用不易。大样本置换检验的目的，就是在样本大小很大时 用一个熟知的连续型分布（一般是正态分布）去逼近这一分布，因而可以容易地决定否定域临界值的近似值。

大样本置换检验的另一个重要意义，是通过它可看出置换检验与传统检验（t 检验、F 检验等）的联系，从而给这些熟知的检验以一种新的解释。

上节的例 5·2 和例 5·3，代表了两种不同的情况，在例 5·2 中，参与置换的变量数目随样本大小 n 增加至于无穷。而在例

5·3 中，参与置换的变量的组数随 n 增加，但每组内参与置换的变量数固定不变。在前一场合需要与置换有关的特殊性质的极限定理，在后一场合，则只须用到普通的中心极限定理。

一、线性置换统计量的渐近正态性

定义5·3 给定两组常数 $a_1,\cdots,a_n$ 及 $b_1,\cdots,b_n$。设$\xi=(\xi_1,\cdots,\xi_n)$为 n 维随机向量，其分布是：ξ 以等概率 $1/n!$ 取$(b_1,\cdots,b_n)$的任一置换：

$$P((\xi_1,\cdots,\xi_n)=(b_{i_1},\cdots,b_{i_n}))=1/n!$$ ，对$(1,\cdots,n)$ 的任一置换$(i_1,\cdots i_n)$， （5·15）

则称

$$L=a_1\xi_1+\cdots+a_n\xi_n \qquad (5\cdot16)$$

为一线性置换统计量。

在§4·1 的一段中我们曾定义过线性秩统计量 $L=c_1a(R_1)+\cdots+c_na(R_n)$。不难看到，这与（5·16）在实质上是一回事。此处的 c_i 相当于（5·16）中的 a_i。又若记 $a(i)=b_i$，$i=1,\cdots,n$。则据定理 4·1，n 维随机向量$(a(R_1),\cdots,a(R_n))$取$(b_1,\cdots,b_n)$的任一置换的概率为 $1/n!$ ，因而 $\sum\limits_{i=1}^{n}c_ia(R_i)$的分布与（5·16）的分布相同。从统计应用的角度看，二者的差别在于，在线性秩统计量中$a(\cdot)$一般是一有规则的计分函数，而在线性置换统计量中，$(b_1,\cdots,b_n)$往往是随机变量所取的值，性质较复杂些。

据上述（5·16）与线性秩统计量的关系，由公式（4·3）和（4·4），即得（5·16）定义的 L 的期望和方差为：

$$E(L)=n\bar{a}\bar{b},\quad \mathrm{Var}(L)=\frac{1}{n-1}\sum_{i=1}^{n}(a_i-\bar{a})^2\sum_{i=1}^{n}(b_i-\bar{b})^2, \qquad (5\cdot17)$$

其中$\bar{a}=\sum\limits_{i=1}^{n}a_i/n$，$\bar{b}=\sum\limits_{i=1}^{n}b_i/n$。

例5·6 考察例 5·2 中的统计量 T 在给定统计量 $M(Z)$的条件下的条件分布．按定义，这分布相当于

$$L=-\frac{1}{n_1}\xi_1-\frac{1}{n_1}\xi_2-\cdots-\frac{1}{n_1}\xi_{n_1}+\frac{1}{n_2}\xi_{n_1+1}+\cdots+\frac{1}{n_2}\xi_n \tag{5·18}$$

的分布，其中（$\xi_1,\cdots,\xi_n$）以等概率 $1/n!$ 取（$Z_1,\cdots,Z_n$）的任一置换。这相当于（5·16）中

$$(a_1,\cdots,a_n)=\left(-\frac{1}{n_1},\cdots,-\frac{1}{n_1},\ \frac{1}{n_2},\ \cdots,\ \frac{1}{n_2}\right) \tag{5·19}$$

$$(b_1,\cdots,b_n)=(Z_1,\cdots,Z_n) \tag{5·20}$$

的情形。回忆$(Z_1,\cdots,Z_n)=(X_1,\cdots,X_{n_1},\ Y_1,\cdots,Y_{n_2})$。当然，这一切都是在原假设成立的前提下。

注意在本例中，其实也是在一切置换检验中，$(b_1,\cdots,b_n)$ 其实是随机变量的取值。但我们是在给定 $Z_1,\cdots,Z_n$ 的条件下去讨论的，它们都当作常数看待。

现在把定义 5·3 中的 $a_1,\cdots,a_n$ 改为$a_{n1},\cdots,a_{nn}$，$b_1,\cdots,b_n$ 改为$b_{n1},\cdots,b_{nn}$，$\xi_1,\cdots,\xi_n$ 改为 $\xi_{n1},\cdots,\xi_{nn}$，而将所定义的线性置换统计量（5·16）改记为 L_n。这个修改的意义是：我们要考虑一串线性置换统计量$\{L_n\}$，且在定义 L_n 时，所用到的常数都从新来过，与定义 L_{n-1}时用过的常数无关。把 $a_{n1},\cdots,a_{nn}$ 和$b_{n1},\cdots,b_{nn}$ 的平均分别记为 $\bar{a}_n$ 和$\bar{b}_n$，L_n 的期望和方差分别记为 l_n 和 σ_n^2。我们要考虑的问题是：在何种条件下，标准化变量 $(L_n-l_n)/\sigma_n$ 当 $n\to\infty$ 时依分布收敛于 $N(0,1)$。为此，要定义一个由 Wald 和 Wolfowitz 在 1944 年引进的条件：

定义5·4(条件WW)． 若对任何固定的自然数 $r\geqslant 3$，序列

$$\left\{\frac{1}{n}\sum_{i=1}^{n}(a_{ni}-\bar{a}_n)^r\Big/\left(\frac{1}{n}\sum_{i=1}^{n}(a_{ni}-\bar{a}_n)^2\right)^{r/2}:\ n=2,3,4,\cdots\right\} \tag{5·21}$$

保持有界(其界可与 r 有关)，则称序列

$$\{(a_{n1},\cdots,a_{nn}):n=1,2,\cdots\} \tag{5·22}$$

满足条件 WW．

在定义 4·2 中曾引进形如（5·22）的序列满足条件 N 的概

念.不难看出：条件 WW 的要求比条件 N 高.事实上，若(5·22)不满足条件 N，则它绝不可能满足条件 WW.事实上，因(5·22)不满足 N，故存在 $\varepsilon>0$ 及一串上升的自然数$\{n_k\}$，使

$$\max_{1\leq i\leq n_k}(a_{n_k i}-\bar{a}_{n_k})^2\geqslant\varepsilon\sum_{i=1}^{n_k}(a_{n_k i}-\bar{a}_{n_k})^2,\ k=1,2,\cdots$$

于是有

$$\frac{1}{n_k}\sum_{i=1}^{n_k}(a_{n_k i}-\bar{a}_{n_k})^4\geqslant\frac{1}{n_k}\Big(\max_{1\leq i\leq n_k}(a_{n_k i}-\bar{a}_{n_k})^2\Big)^2$$

$$\geqslant\frac{\varepsilon^2}{n_k}\Big(\sum_{i=1}^{n_k}(a_{n_k i}-\bar{a}_{n_k})^2\Big)^2=\varepsilon^2 n_k\Big(\frac{1}{n_k}\sum_{i=1}^{n_k}(a_{n_k i}-\bar{a}_{n_k})^2\Big)^2,$$

对 $k=1,2,\cdots$ 成立. 由于 $\varepsilon>0$ 而 $n_k\to\infty$，知序列(5·21)在 $r=4$ 时已不有界，故条件 WW 不能成立.

现在我们可以陈述下面的定理：

定理5·1 如果两序列$\{(a_{n1},\cdots,a_{nn}):n=1,2,\cdots\}$及$\{(b_{n1},\cdots,b_{nn}):n=1,2,\cdots\}$中，有一个满足条件 WW 而另一个满足条件 N，则由它们所定义的线性置换统计量序列$\{L_n\}$满足

$$(L_n-l_n)/\sigma_n\xrightarrow{\mathscr{L}}N(0,1),\ \text{当}\ n\to\infty.\tag{5·23}$$

这个定理的原型是 Wald 和 Wolfowitz 在 1944 年提出，当时他们要求两序列都满足 WW. 后来 Noether 在 1949 年改进为上述形式. 好奇的读者或许会问：既然如此，能否进一步改进为只要求两序列都满足条件 N？这是不可能的. 事实上，本书作者之一曾在一项工作中证明：若只假定两序列满足条件 N，则 $(L_n-l_n)/\sigma_n$ 可以没有极限分布，也可以依分布收敛于任一个方差有界的无穷可分分布. 但是，Hajek 在 1961 年得出了一个深刻结果. 他证明了：若两序列都满足条件 N，则 $(L_n-l_n)/\sigma_n\xrightarrow{\mathscr{L}}N(0,1)$ 的充要条件是，这两序列还满足由Motto 在 1955年引进的一个条件 M. 这也就是我们曾在§4·1 的二段开头处提到的 Hajek 结果.

我们将不给出 Hajek 定理的陈述和证明. 有兴趣的读者可参看在 §4·1 的二段处引的 Hajek 文章，也可参看陈希孺《数理

统计引论》P.638-645. 我们将给出定理 5·1 的证明.但为不打断此处的叙述，这证明将写在文章附录内。

顺便提到，由于前述的线性置换统计量与线性秩统计量在分布上的同一性，定理 5·1 也可用于证明定理 4·4 的某些特例。这特例包含了多数在应用上重要的情况. 这也将在附录中给出。

在多样本问题中要同时考虑若干个线性置换统计量的联合分布。设有 $m+1$ 个序列

$\{(a_{n1}^{(k)},\cdots,a_{nn}^{(k)}):n=1,2,\cdots\},\ k=1,\cdots,m;$

$\{(b_{n1},\cdots,b_{nn}):n=1,2,\cdots\},$

把由 $(a_{n1}^{(k)},\cdots,a_{nn}^{(k)})$ 和 $(b_{n1},\cdots,b_{nn})$ 产生的线性置换统计量记为 L_{nk}，其数学期望与方差分别记为 l_{nk} 和 σ_{nk}^2。记

$$L_n=\left(\frac{L_{n1}-l_{n1}}{\sigma_{n1}},\ \cdots,\ \frac{L_{nm}-l_{nm}}{\sigma_{nm}}\right)'.$$

定理 5·2 若对每个 $k=1,\cdots,m$,序列 $\{(a_{n1}^{(k)},\cdots,a_{nn}^{(k)}):n=1,2,\cdots\}$ 适合条件 WW，而 $\{(b_{n1},\cdots,b_{nn}):n=1,2,\cdots\}$，适合条件 N，又设极限

$$\lim_{n\to\infty}\frac{\sum_{i=1}^{n}(a_{ni}^{(u)}-\bar{a}_n^{(u)})(a_{ni}^{(v)}-\bar{a}_n^{(v)})}{\left(\sum_{i=1}^{n}(a_{ni}^{(u)}-\bar{a}_n^{(u)})^2\sum_{i=1}^{n}(a_{ni}^{(v)}-\bar{a}_n^{(v)})^2\right)^{1/2}}=\lambda_{uv}, \tag{5·24}$$

对任何 $u\neq v$，u，$v=1,\cdots,m$ 存在(当 $u=v$ 时，极限当然存在且 $\lambda_{uu}=1$)，且 $\Lambda=(\lambda_{uv})_{u,v=1,\cdots m}$ 为满秩方阵，则当 $n\to\infty$ 时，有

$$L_n\xrightarrow{\mathscr{L}}N(0,\Lambda). \tag{5·25}$$

本定理不难在定理 5·1 的基础上去证明，细节也不在此给出了。

二、前段结果的应用

现在我们要将(一)中的极限定理用于 §5·1 的几个例子中，以决定置换检验否定域临界值的大样本近似。为此需要验证某些序列满足条件 WW 或者 N.我们把需要的结果列举并证明于下：

(1) 设$(a_{n1},\cdots,a_{nn})=(d_n,\cdots,d_n,e_n,\cdots,e_n)$, $n=1,2,\cdots$。其中d_n有n_1个，e_n有n_2个，$n_1+n_2=n$，且$d_n\neq e_n$，则如存在$\lambda>0$使对一切n有$\lambda\leqslant n_1/n\leqslant 1-\lambda$，则$\{(a_{n1},\cdots,a_{nn}):n=1,2,\cdots\}$满足条件WW。

证明可由简单计算直接得到，留给读者.

(2) 设$n_1,\cdots,n_c$为自然数，$n_1+\cdots+n_c=n$。令

$$(a_{n1}^{(k)},\cdots,a_{nn}^{(k)})=(0,\cdots0,\ \cdots,0,\cdots,0,1,\cdots,1,0,\cdots,0,\cdots,0,\cdots,0) \qquad (5\cdot26)$$

$$k=1,2,\cdots c$$

在上述向量中，全部坐标划分为c段，当$i\neq k$时，第i段含n_i个0，而第k段则含n_k个1．假定存在$\lambda>0$，使对一切n和$k=1,\cdots,c$有

$$n_k/n\geqslant\lambda \qquad (5\cdot27)$$

则对每一个k，$k=1,\cdots,c$，$\{(a_{n1}^{(k)},\cdots,a_{nn}^{(k)}):n=1,2,\cdots\}$适合条件WW。又若对任何$k=1,\cdots,c$，极限

$$\lim_{n\to\infty}n_k/n=\rho_k\text{存在且}>0, \qquad (5\cdot28)$$

则极限(5·24)存在，且

$$\lambda_{uv}=-\left(\frac{\rho_u\rho_v}{(1-\rho_u)(1-\rho_v)}\right)^{1/2}\text{当}u\neq v\ (\lambda_{uu}=1). \qquad (5\cdot29)$$

前一结论包含在(1)中，后一结论易由直接计算得到，留给读者.

(3) 设$X_1,X_2,\cdots$为一串独立同分布的随机变量，对任何自然数r，有$E|X_1|^r<\infty$，又$\mathrm{Var}(X_1)>0$（注意由$EX_1^2<\infty$有$\mathrm{Var}(X_1)<\infty$）。令

$$(a_{n1},\cdots,a_{nn})=(X_1,\cdots,X_n),n=1,2,\cdots, \qquad (5\cdot30)$$

则以概率1成立：这个序列满足条件WW。

证明　以μ_r记X_1的r阶中心矩：

$$\mu_r=E(X_1-EX_1)^r,\quad r=2,3,4,\cdots$$

则由Колмогоров强大数律易知以概率1有

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X}_n)^r=\mu_r\left(\overline{X}_n=\sum_{i=1}^{n}X_i/n\right),$$

由于$\mu_2>0$，知以概率1有

$$\lim_{n\to\infty}\left\{\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X}_n)^r\right)\Big/\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X}_n)^2\right)^{r/2}\right\}=\mu_r/\mu_2^{r/2}.$$

这证明了：以概率1，上式左边极限号下的序列为有界，因而证明了所要的结果。

(4) 设$X_1,X_2,\cdots$为一串独立同分布的随机变量，$0<\mathrm{Var}(X_1)<\infty$，定义$(a_{n1},\cdots,a_{nn})$如(5·30)。则以概率1成立：这序列满足条件$N$。

证明　按条件N的定义，要证明：

$$\lim_{n\to\infty}\left\{\frac{\frac{1}{n}\max\limits_{1\leqslant i\leqslant n}(X_i-\overline{X}_n)^2}{\frac{1}{n}\sum\limits_{i=1}^{n}(X_i-\overline{X}_n)^2}\right\}=0,\ \text{a.s.},\tag{5·31}$$

因为$0<\mathrm{Var}(X_1)<\infty$，按强大数律，上式极限号下表达式的分母以概率1收敛于$\mathrm{Var}(X_1)>0$。由此可知，(5·31)等价于

$$\lim_{n\to\infty}\frac{1}{n}\max_{1\leqslant i\leqslant n}(X_i-\overline{X}_n)^2=0,\ \text{a.s.}\tag{5·32}$$

若以$X_{(1)}\leqslant\cdots\leqslant X_{(n)}$记$X_1,\cdots,X_n$的次序统计量，则易见

$$\max_{1\leqslant i\leqslant n}(X_i-\overline{X}_n)^2\leqslant(X_{(n)}-X_{(1)})^2\leqslant2(X_{(n)}^2+X_{(1)}^2)\leqslant4\max_{1\leqslant i\leqslant n}X_i^2\tag{5·33}$$

现往证下面的事实：若$\xi_1,\xi_2,\cdots$独立同分布，$E|\xi_1|<\infty$，则$\lim\limits_{n\to\infty}\frac{1}{n}\max\limits_{1\leqslant i\leqslant n}|\xi_i|=0$, a.s.. 为证此，任给定$\varepsilon>0$。定义一列事件

$$A_i=\{|\xi_i|>i\varepsilon\},\quad i=1,2,\cdots,\tag{5·34}$$

以F记$|\xi_1|$的分布函数，有

$$P(A_i)=1-F(i\varepsilon)=\sum_{k=i}^{\infty}(F((k+1)\varepsilon)-F(k\varepsilon)).$$

于是由$E|\xi_1|<\infty$知

$$\sum_{i=1}^{\infty}P(A_i)=\sum_{i=1}^{\infty}\sum_{k=i}^{\infty}\{F((k+1)\varepsilon)-F(k\varepsilon)\}$$

$$=\sum_{i=1}^{\infty} i\{F((i+1)\varepsilon)-F(i\varepsilon)\}$$

$$=\sum_{i=1}^{\infty} i\varepsilon P(i\varepsilon<|\xi_1|\leqslant(i+1)\varepsilon)/\varepsilon\leqslant E|\xi_1|/\varepsilon<\infty$$

故由 Borel-Cantelli 引理，知

$$P\{事件列 A_1,A_2,\cdots只发生有限个\}=1 \quad (5\cdot35)$$

但"事件 $A_1,A_2,\cdots$只发生有限个"意味着当n充分大时，有A_n不发生，即当n充分大时有$|\xi_n|\leqslant n\varepsilon$。由此不难推出：当$n$充分大时有$\max\limits_{1\leqslant i\leqslant n}|\xi_i|\leqslant n\varepsilon$，这进一步得出$\limsup\limits_{n\to\infty}\frac{1}{n}\max\limits_{1\leqslant i\leqslant n}|\xi_i|\leqslant\varepsilon$。因此，由（5·35）推出：对任给$\varepsilon>0$，以概率1成立

$$\limsup_{n\to\infty}\frac{1}{n}\max_{1\leqslant i\leqslant n}|\xi_i|\leqslant\varepsilon$$

由于$\varepsilon>0$的任意性，这证明了$\lim\limits_{n\to\infty}\frac{1}{n}\max\limits_{1\leqslant i\leqslant n}|\xi_i|=0$，a.s.。注意到$\mathrm{Var}(X_1)<\infty$因而$EX_1^2<\infty$，把这一结论用于序列$(\xi_1,\xi_2,\cdots)=(X_1^2,X_2^2,\cdots)$，得到

$$\lim_{n\to\infty}\frac{1}{n}\max_{1\leqslant i\leqslant n}X_i^2=0,\text{a.s.} \quad (5\cdot36)$$

由（5·33）和（5·36）即证明了（5·32），因而证明了（5·31）。

有了这些准备，我们可以继续§5·1中诸例的讨论。

例5·2′ 沿用例5.2的符号，我们来讨论其大样本置换检验的临界值的确定。除了我们曾作的假定外，还要作两个假定。

（1）存在$\lambda>0$，使对一切n，有$\lambda\leqslant n_1/n\leqslant1-\lambda$。

（2）总体分布F的方差非0有限①。

按照我们的记号，$(Z_1,\cdots,Z_n)=(X_1,\cdots,X_{n_1},Y_1,\cdots,Y_{n_2})$。暂记$\bar{Z}_1=\sum\limits_{i=1}^{n_1}Z_i/n_1$，$\bar{Z}_2=\sum\limits_{i=n_1+1}^{n}Z_i/n_2$，统计量$T(Z)$就是$\bar{Z}_2-\bar{Z}_1$。在例5·6中已指出：在原假设成立时，在给定$M(Z)$之下，$T$的条

① 这里是在例5·2的模型1之下去讨论，若用模型2，则须直接假定$(X_1,\cdots,X_n)$和$(Y_1,\cdots,Y_n)$满足条件N。这时用不着预备事实(4)。

件分布与置换统计量（5·18）同，即由（5·19）和（5·20）决定的置换统计量．根据假定 a，b，及上面的预备事实 1° 和 4°，知定理 5·1 的条件满足（确切地说，应为"以概率 1 满足"。下两例同此），又易算出 $l_n=0$，及

$$\begin{aligned}\sigma_n^2&=\frac{1}{n-1}\frac{n}{n_1n_2}\sum_{i=1}^{n}(Z_i-\bar{Z})^2\\&=\frac{n}{(n-1)n_1n_2}\left\{\sum_{i=1}^{n_1}(Z_i-\bar{Z}_2)^2+\sum_{i=n_1+1}^{n}(Z_i-\bar{Z}_2)^2\right.\\&\quad\left.+\frac{n_1n_2}{n}(\bar{Z}_1-\bar{Z}_2)^2\right\}\\&=A\{\xi^2+B(\bar{Z}_1-\bar{Z}_2)^2\},\end{aligned}\tag{5·37}$$

其中 $A=n/(n-1)n_1n_2$，$\xi^2=\sum_{i=1}^{n_1}(Z_i-\bar{Z}_1)^2+\sum_{i=n_2+1}^{n}(Z_i-\bar{Z}_2)^2$，$B=n-1$，据定理 5·1，本例原假设的置换检验的否定域，当n甚大时可近似地取为

$$|\bar{Z}_2-\bar{Z}_1|>\sqrt{A(\xi^2+B(\bar{Z}_1-\bar{Z}_2)^2}\,u_{\alpha/2}^2,\tag{5·38}$$

α 为给定的水平．不难看出，当 $n-1>u_{\alpha/2}^2$时，(5·38) 等价于

$$|\bar{Z}_2-\bar{Z}_1|/\xi>\frac{\sqrt{n}}{\sqrt{n-1-u_{\alpha/2}^2}}\frac{1}{\sqrt{n_1n_2}}u_{\alpha/2},\tag{5·39}$$

把记号 Z 改回到 X，Y，并拼凑成通常的两样本 t 统计量

$$t=\sqrt{\frac{n_1n_2(n-2)}{n}}(\bar{Y}-\bar{X})\Big/\sqrt{\sum_{i=1}^{n_1}(X_i-\bar{X})^2+\sum_{i=1}^{n_2}(Y_i-\bar{Y})^2},$$

可将 (5·39) 写为等价的形式：

$$|t|>\sqrt{\frac{n-2}{n-1-u_{\alpha/2}^2}}u_{\alpha/2}.\tag{5·40}$$

如果是单侧假设（对立假设：药物 B 优于 A），则(5·40) 相应地改成单边的形式

$$t>\sqrt{\frac{n-2}{n-1-u_{\alpha}^2}}u_{\alpha},\tag{5·41}$$

这个结果很有意思，因为它形式上与通常的两样本 t 检验完全一样，只是临界值略有不同；但当 n 甚大时，$(n-2)/(n-1-u_{\alpha/2}^2)$

接近 1 而 $t_{n-2}(\alpha/2)$ 接近 $u_{\alpha/2}$. 故二者的临界值也相差很小.

这样，当样本大小 n 很大时，通过使用置换检验再取其大样本逼近,我们基本上又回到了常用的 t 检验.它让人看起来好象是转了一个大圈子而一无所获. 这确是置换检验的一个弱点，也说明了为何它未获得广泛的应用：如果我们坚持按置换检验的原义去做，则除非 n 很小，计算量将极大，即便可行，代价也太大了. 若用极限分布去逼近之，则基本上又回到传统检验. 这个两难之局并无妥善的处理方法.

但我们也不应据此而完全否定置换检验的意义. 其理由实际上在前面已指出过了. 此处要重复强调一下.

1. 在应用上，如在本例的情况，往往 n 并不很大，例如 $n\leqslant 30$ 的情况，在当今的计算条件下，并不算很过分. 而恰好在这个情况下，传统模型中的假定，问题较多. 例如，不论是否正态都采用 u 检验,而以中心极限定理去解释之. 但在 n 不甚大时，统计量的精确分布与正态分布可相去甚远，从而使名义上的水平 α 与事实上的水平有显著差别，但使用者无法知道这差别是多少. 置换检验则能在对总体分布不作任何假定的情况下，提供一个具有确切水平 α 的检验，这是很了不起的.

2. 即使在传统检验也可以使用的场合，置换检验理论给这种检验一个新的解释. 这种解释是建立在更现实的统计假定的基础上. 例如，如前面曾指出的，在传统模型下看不出在试验中施行随机化对尔后的统计分析起了何作用，但在置换检验理论中这一点看得很清楚，事实上，随机化原则是这种理论的柱石.

例5·4′ 现在来考虑多样本问题，并沿用例 5·4 的记号. 在该例中，我们曾提出由 (5·14) 定义的统计量 T，在它的基础上作置换检验.

设原假设成立. 要求统计量 T 在给定 $M(Z)$ 之下的条件分布的极限分布.这里用得着定理 5·2. 取$(a_{n1}^{(k)}\cdots,\ a_{nn}^{(k)})=(0,\cdots 0,\cdots 1/n_k,\ \cdots,1/n_k,\ \cdots,\ 0,\ \cdots,0)$，其中 $1/n_k$ 占据长为 n_k 的一

段而$k=1,\cdots,c-1$. 又$(b_1,\cdots,b_n)=(Z_1,\cdots,Z_n)$. 把由这两个序列定义的置换统计量记为$L_{nk}$，则

$$L_{nk}=\sum_{i=n_1+\cdots+n_{k-1}+1}^{n_1+\cdots+n_k} Z_i/n_k=\bar{Z}_k,\ k=1,\cdots,c-1.$$

按公式(5·17)，易算得

$$l_{nk}=E(L_{nk})=n_k\bar{Z},$$

$$\sigma_{nk}^2=\mathrm{Var}(L_{nk})=\frac{1}{n-1}\frac{n_k(n-n_k)}{n}\sum_{i=1}^{n}(Z_i-\bar{Z})^2,$$

以下我们假定(5·28)成立，且总体分布F的方差非0有限（参看例5·2′的足注）。这时，据预备事实1°，4°及定理5·2，以概率1，当$n\to\infty$时，$c-1$维随机向量$L_n=\left(\frac{L_{n1}-l_{n1}}{\sigma_{n1}},\ \cdots\frac{L_{n,c-1}-l_{n,c-1}}{\sigma_{n,c-1}}\right)'$在给定$M(Z)$之下，依分布收敛于$c-1$维正态分布$N(0,\ \Lambda)$，其中

$$\Lambda=(\lambda_{uv})_{u,v=1,\cdots,c-1},\quad \lambda_{uv}\text{ 由 }(5\cdot 29)\text{确定。}$$

容易证明(请读者自证)，Λ为非异正定方阵。故有

$$L_n'\Lambda^{-1}L_n\xrightarrow{\mathscr{L}}\chi^2_{c-1},\tag{5·42}$$

定义$\Lambda_n=(\lambda_{nuv})_{u,v=1,\cdots,c-1}$，其中

$$\lambda_{nuu}=1,\lambda_{nuv}=-\sqrt{\frac{n_un_v}{(n-n_u)(n-n_v)}},\quad u\neq v$$

则$\Lambda_n\to\Lambda$当$n\to\infty$，因而(5·42)可改为

$$L_n'\Lambda_n^{-1}L_n\xrightarrow{\mathscr{L}}\chi^2_{c-1},\tag{5·43}$$

计算$L_n'\Lambda_n^{-1}L_n$.直接验证可知

$$\Lambda_n^{-1}=\mathrm{diag}\left(\frac{n-n_1}{n},\ \cdots,\ \frac{n-n_{c-1}}{n}\right)+vv'\frac{n}{n_c}$$

$$(v=(\sqrt{n_1(n-n_1)}/n,\ \cdots,\ \sqrt{n_{c-1}(n-n_{c-1})}/n)'$$

经过少量的化简，得到

$$L_n'\Lambda_n^{-1}L_n=(n-1)\sum_{i=1}^{c-1}n_i(\bar{Z}_i-\bar{Z})^2/S$$

$$+(n-1)\frac{1}{n_c}\left(\sum_{i=1}^{c-1}n_i(Z_i-\bar{Z})\right)^2/S,$$

其中 $S=\sum_{i=1}^{n}(Z_i-\bar{Z})^2$. 由于 $\sum_{i=1}^{c}n_i(\bar{Z}_i-\bar{Z})=0$，知 $\sum_{i=1}^{c-1}n_i(\bar{Z}_i-\bar{Z})=-n_c(\bar{Z}_c-\bar{Z})$. 于是得

$$L_n'\Lambda_n^{-1}L_n=(n-1)\sum_{i=1}^{c}n_i(\bar{Z}_i-\bar{Z})^2/S=(n-1)T/S.$$

T 由(5·14)定义。总结上述，我们证明了：在所设的条件下，以概率为 1 地，当给定$M(Z)$时，有$(n-1)T/S\xrightarrow{\mathscr{L}}\chi_{c-1}^2$.

因此，本例原假设的置换检验的否定域，当 n 甚大时，近似地可取为

$$(n-1)T/S>\chi_{c-1}^2(\alpha) \tag{5·44}$$

把 $Z_1,Z_2,\cdots,Z_n$ 还原成 $X_{11},\cdots,X_{1n_1},\cdots,X_{c1},\cdots,X_{cn_c}$，注意到

$$S=\sum_{i=1}^{c}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_i)^2+T.$$

易见(5·44)等价于

$$\frac{\frac{1}{c-1}\sum_{i=1}^{c}n_i(\bar{X}_i-\bar{X})^2}{\frac{1}{n-c}\sum_{i=1}^{c}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_i)^2}>\frac{(n-c)\chi_{c-1}^2(\alpha)}{(c-1)(n-1-\chi_{c-1}^2(\alpha))} \tag{5·45}$$

上式左边的统计量不是别的，正是一因素方差分析中的 F统计量。故我们基本上又回到了传统的 F 检验，只是临界值由$F_{c-1,n-c}(\alpha)$改变为(5·45)的右边。当然，此处未假定总体分布为正态。还须注意，当 $n\to\infty$ 时，$F_{c-1,n-c}(\alpha)$及(5·45)的右边。都以 $\chi_{c-1}^2(\alpha)/(c-1)$ 为极限，故当 n 很大时二者很接近。

这个结果对置换检验的意义，以及有关的解释，与例5·2′相似。此处不赘述了。

例5·5′ 沿用例 5·5 的记号。设 X,Y 两个随机变量都有非0 有限的方差，且其中有一个，例如X，有任意阶的有限矩①。这时，根据预备事实(3)和(4)，以概率1成立下述事实：$\{(X_1,\cdots,X_n):n=1,2,\cdots\}$ 满足条件 WW 而$\{(Y_1,\cdots,Y_n):n=1,2,\cdots\}$满足条件 N. 又在例 5·5 中已说明过，若原假设"X，Y独立"成

① 使用Hajek 更深刻的结果可证明，r阶矩有限的条件可免除。

立 则在给定 $M=(M_1, M_2)$（见例5·5）之下，$\sum\limits_{i=1}^{n}X_iY_i$的条件分布即是由

$$(a_1,\cdots,a_n)=(X_1,\cdots,X_n),(b_1,\cdots,b_n)=(Y_1,\cdots,Y_n)$$

按定义 5·3 而产生的线性置换统计量。按公式(5·17)，此置换统计量的期望和方差分别为 $l_n=n\bar{X}\bar{Y}$ 及 $\sigma_n^2=\frac{1}{n-1}\sum\limits_{i=1}^{n}(X_i-\bar{X})^2\sum\limits_{i=1}^{n}(Y_i-\bar{Y})^2$. 因此据定理 5·1 有

$$\sqrt{n-1}\frac{\sum\limits_{i=1}^{n}X_iY_i-n\bar{X}\bar{Y}}{\left(\sum\limits_{i=1}^{n}(X_i-\bar{X})^2\sum\limits_{i=1}^{n}(Y_i-\bar{Y})^2\right)^{1/2}}\xrightarrow{\mathscr{L}}N(0,1) \quad (5\cdot46)$$

注意(5.46)不是在无条件的意义上，而是在给定 $M=(M_1,M_2)$ 的条件下，且这事实也只是以概率 1 成立，引进样本相关系数r，则据(5·46)，原假设的水平 α 的大样本否定域可取为

$$|r|>u_{\alpha/2}/\sqrt{n-1}. \quad (5\cdot47)$$

如果假定总体分布为二维正态分布，则可得到水平 α 的确切检验。如所周知，这检验有否定域

$$|r|>t_{n-2}\left(\frac{\alpha}{2}\right)\Big/\sqrt{n-2+t_{n-2}^2\left(\frac{\alpha}{2}\right)}. \quad (5\cdot48)$$

比较(5·47)和(5·48)，我们又看到了前面两例中的现象：大样本置换检验与传统检验一样，都是以样本相关系数绝对值的大值为否定域，但界限有些不同，但是，当 $n\to\infty$ 时$t_{n-2}\left(\frac{\alpha}{2}\right)\to u_{\alpha/2}$ 而 $\sqrt{n-1}\Big/\sqrt{n-2+t_{n-2}^2\left(\frac{\alpha}{2}\right)}\to 1$，故当 n 很大时，这二者的界限，只相差一个数量级为 $o\left(\frac{1}{\sqrt{n}}\right)$的无穷小量。

三、另一种例子：随机区组

设有 c 个种子品种，在 n 个区组内做试验。每个区组内包含 c 个"小区"，恰够每个品种各做一次。

我们并不假定每区组内各小区为绝对均匀。相反，小区之间的差异正好是产生随机误差的来源，且构成统计模型的依据。

我们假定：第 j 区组的第 k 小区有一个反映该小区的条件的常数 a_{jk}，$j=1,\cdots,n$，$k=1,\cdots,c$，而第 i 个品种则有一个反映该品种优良性的常数 θ_i，$i=1,\cdots,c$.如果把品种 i 种在第 j 区组的 k 小区上，其亩产将为 θ_i+a_{jk}. "品种无差异"的原假设可表为"$\theta_1=\cdots=\theta_c$"。

根据随机区组试验的安排，在每个区组内，把各小区随机地配给于 c 个品种，且在这 n 个区组内，随机化是独立进行的。这样，若以 X_{ij} 记第 i 品种在第j区组内的（分配给它那小区上的）亩产，则将有模型

$$(X_{1j},\cdots,X_{cj})'=(\theta_1,\cdots,\theta_c)'+\xi_j,\quad j=1,2,\cdots,n \tag{5·49}$$

此处 $\xi_1,\xi_2,\cdots,\xi_n$ 相互独立，且 ξ_j' 取 $(a_{j1},\cdots,a_{jc})$ 的任一置换的概率都是 $1/c!$.

记 $\bar{X}_i=\sum_{j=1}^{n}X_{ij}/n$，$\bar{X}=\sum_{i=1}^{c}\sum_{j=1}^{n}X_{ij}/nc$. 以

$$T=\sum_{i=1}^{c}(\bar{X}_i-\bar{X})^2 \tag{5·50}$$

作为衡量试验结果与原假设的偏差的指标，其理由是：当 $\theta_1=\cdots=\theta_c$ 成立时，$\bar{X}_1,\cdots,\bar{X}_c$ 有相同的期望，它们应比较接近，这导致 T 取小值. 反之，则 T 将倾向于取大值。故可以取

$$T>C \tag{5·51}$$

作为否定域。为定C,要定出T在原假设下的分布。当原假设成立时,每个 ξ_j 独立地各可取 $C!$ 个值,且都以 $1/c!$ 的等概率。由此可知,T以等概率取$(c!)^n$个值。这些值是按下面的方法算出的:对每个 j, 把 $X_{1j},\cdots,X_{cj}$ 任意置换成一个新的次序,把置换后所得的结果当作 $X_{1j},\cdots,X_{cj}$, $j=1,\cdots,n$, 按公式(5·50)就算出一个 T 值. 取一切可能的置换都作这个计算,则得出$(c!)^n$ 个T值①。

---

①从对称性考虑可知，可以把第一区组的c个值固定不予置换。故实际上不同之值只有$(c!)^{n-1}$个。例如3品种4区组有$(3!)^3=216$个不同的T值。

按其大小排列为 $t_1 \geqslant t_2 \geqslant \cdots \geqslant t_N, N=(c!)^n$. 用原有的数据 $(X_{1j}, \cdots, X_{cj})$ (不作置换)算出的 T 值仍记为 T. 若 $T \geqslant t_{N\alpha}$, 则否定原假设, 不然就接受原假设.

这种检验仍属置换检验, 不过可允许的置换受到限制: 只能在一区组内作置换, 而不能把不同区组内的值彼此置换. 这检验无须通过另一统计量进行条件化, 故并非条件检验.

以上讨论的模型相当于例 5·2 的模型 2. 也可以引进类似于例 5·2 模型 1 的模型, 这就是在 §4·3 的二段中已讨论过的模型

$$X_{ij}=\mu+a_i+b_j+e_{ij}, i=1,\cdots,c, j=1,\cdots,n \tag{5·52}$$

a_i, b_i 分别反映品种组区组效应, e_{ij} 为随机误差. 设 nc 个 e_{ij} 独立同分布, 其公共分布 F 未知. 以 M_j 记 $(X_{1j}, \cdots, X_{cj})$ 的次序统计量, 而 $M=(M_1, \cdots, M_n)$. T 的定义仍如(5·50). 则容易看出: 在给定 M 的条件下, T 的条件分布正与我们刚才描述的一样, 因而可以用完全一样的方式去检验原假设"品种无差异". 当然, 这二者有本质的不同. 即一个是无条件检验而另一个是条件检验, 其中统计量 T 的分布的意义不同.

当 c 或 n 不很小时, 直接通过计算 T 的一切值去进行检验, 计算量太大, 以下我们考虑 T (在原假设下)的极限分布. 如上面指出的, 有两种不同的模型——其一非条件化而另一为条件化, 故以下讨论的极限分布, 也可以是通常的极限分布, 或是在给定 M 时 T 的条件分布的极限分布, 视采用哪一模型而定, 但其最后形式并无差别. 为确定计, 以下我们就本段开头描述的模型来讨论.

记 $\xi_j=(X_{1j}, \cdots, X_{c-1,j})'$, $j=1, \cdots, n$. 据本模型假定, $\xi_1, \cdots, \xi_n$ 独立, 而在原假设成立时, $(X_{1j}, \cdots, X_{cj})$ 以概率 $1/c!$ 取 $(a_{j1}, \cdots, a_{jc})$ 的任一置换. 由此, 经过简单计算, 不难得到

$$E(\xi_j)=(a_j, \cdots, a_j)', a_j=\sum_{k=1}^{c} a_{jk}/c,$$

$$\mathrm{Cov}(\xi_j)=S_j^2 A, \quad S_j^2=\frac{1}{c}\sum_{k=1}^{c}(a_{jk}-a_j)^2,$$

$$\Lambda=(\lambda_{ij})_{i,j=1,\cdots,c-1},\lambda_{ii}=1,\lambda_{ij}=-\frac{1}{c-1},\ i\neq j$$

记$\overline{X}=\sum_{i=1}^{c}\sum_{j=1}^{n}X_{ij}/nc$. 则$\sum_{j=1}^{n}(\xi_j-E\xi_j)=n(\overline{X}_1-\overline{X},\cdots,\overline{X}_{c-1}-\overline{X})'$. 记$\tilde{S}_n^2=S_1^2+\cdots+S_n^2$，则在一定条件下可以证明，当$n\to\infty$时有

$$\tilde{S}_n^{-1}\Lambda^{-1/2}\sum_{j=1}^{n}(\xi_j-E\xi_j)\xrightarrow{\mathscr{L}}N(0,I) \tag{5·53}$$

其中I为$c-1$阶单位阵。例如，在下面两个条件之下，

(1) $\{a_{jk}:j=1,2,\cdots,k=1,\cdots,c\}$有界；

(2) $\tilde{S}_n\longrightarrow\infty$ 当$n\to\infty$,

可以证明(5·53)。这两个条件从实用的观点看都很自然而合理。我们把这个结论的证明放在本章附录中，以免打断此处的叙述。

由(5·53)得出

$$n^2(\overline{X}_1-\overline{X},\cdots,\overline{X}_{c-1}-\overline{X})\tilde{S}_n^{-2}\Lambda^{-1}(\overline{X}_1-\overline{X},\cdots,\overline{X}_{c-1}-\overline{X})'$$
$$\xrightarrow{\mathscr{L}}\chi_{c-1}^2, \tag{5·54}$$

直接验证易知Λ^{-1}的(i,i)元为$2(c-1)/c$，而(i,j)元为$(c-1)/c$，当$i\neq j$. 于是不难算出(5·54)的左边（在计算中用到$\sum_{i=1}^{c}(\overline{X}_i-\overline{X})=0$）为

$$(c-1)n^2T\Big/\sum_{i=1}^{c}\sum_{j=1}^{n}(X_{ij}-X_{\cdot j})^2\xrightarrow{\mathscr{L}}\chi_{c-1}^2, \tag{5·55}$$

这里$X_{\cdot j}=\sum_{i=1}^{c}X_{ij}/c$，即上文的$a_j=\sum_{k=1}^{c}a_{jk}/c$，而$T$由(5·50)定义。从(5·50)看出，当$n$充分大时，原假设置换检验的否定域，近似地可取为

$$T>(c-1)^{-1}n^{-2}\sum_{i=1}^{c}\sum_{j=1}^{n}(X_{ij}-X_{\cdot j})^2\chi_{c-1}^2(\alpha), \tag{5·56}$$

记$U=\sum_{i=1}^{c}\sum_{j=1}^{n}(X_{ij}-\overline{X}_i-X_{\cdot j}+\overline{X})^2$，则易证

$$\sum_{i=1}^{c}\sum_{j=1}^{n}(X_{ij}-X_{\cdot j})^2=U+nT,$$

于是(5·56)可改写为

$$\frac{\frac{1}{c-1}nT}{\frac{1}{(c-1)(n-1)}U} > \frac{n-1}{(c-1)n-\chi^2_{c-1}(\alpha)}\chi^2_{c-1}(\alpha). \qquad (5\cdot57)$$

但左边的统计量，就是在通常随机区组设计的方差分析中，为检验"因素效应为 0" 时的 F 统计量，只是在通常方差分析中，右边的界限为 $F_{c-1,(c-1)(n-1)}(\alpha)$. 当$n\to\infty$ 时，这个量，以及(5·57)的右边，都趋向于$\frac{1}{c-1}\chi^2_{c-1}(\alpha)$. 因此，当 n 很大时。本问题的大样本置换检验与传统的 F 检验很接近。

我们这里没有讨论置换检验的大样本功效问题。关于此问题可参看陈希孺《数理统计引论》§6·5 的一段。但我们至少可以看到：在前面讨论的几个重要例子中，置换检验与传统检验有相当的大样本功效。

附：

一、定理 5·1 的证明

定理 5·1 的证明，在文献中见到的有两种方法，一种与定理 2·2 和 3·1 的证明方法相似，是通过从其中分出一个独立和，剩下的余项证明为当 $n\to\infty$ 时依概率收敛于 0，而前者则用通常的中心极限定理去处理。但此法用于置换统计量甚为复杂，故我们在下文将介绍另一种方法。此法虽简单，但也要基于概率论上一个著名的结果，此结果用于正态分布的情况有如下述。

标准正态分布 $N(0,1)$的 r 阶矩为：当 r 为奇数时为 0，r 为偶数时为 $(r-1)!!=1\cdot3\cdot5\cdots(r-1)$. 设 $\{\xi_n\}$ 为一串随机变量，其各阶矩存在有限。以 μ_{nr} 记 $E(\xi_n^r)$. 若

$$\lim_{n\to\infty}\mu_{nr}=\begin{cases}0, & \text{当 } r \text{ 为奇数；}\\ (r-1)!! & \text{当 } r \text{ 为偶数；}\end{cases} \qquad (5A\cdot1)$$

则当 $n\to\infty$ 时，ξ_n 依分布收敛于 $N(0,1)$。

现以L_n记由$(a_{n1},\cdots,a_{nn})$和$(b_{n1},\cdots,b_{nn})$决定的线性置换统计量（定义5·3），l_n和σ_n^2为其数学期望及方差。因为对任何常数c_1，c_2和$d_1\neq0,d_2\neq0$，把a_{ni}改为$d_1a_{ni}+c_1$并把b_{ni}改为$d_2b_{ni}+c_2$后，不改变$(L_n-l_n)/\sigma_n$（这一点很易验证，留给读者）。故如记

$$A_{rn}\overset{令}{=}\sum_{i=1}^n a_{ni}^r,\quad B_{rn}=\sum_{i=1}^n b_{ni}^r, r,\ n=1,\ 2,\ \cdots,$$

则不失普遍性可设

$$A_{1n}=B_{1n}=0,\quad A_{2n}=B_{2n}=n. \tag{5A·2}$$

据此，由$\{(a_{n1},\cdots,a_{nn}):n=1,2,\cdots\}$和$\{(b_{n1},\cdots,b_{nn}):n=1,2,\cdots\}$分别满足条件WW和N，可知

$$A_{rn}=O(n),\quad B_{rn}=o(n^{r/2}),\quad r=3,\ 4,\ \cdots \tag{5A·3}$$

前一结论是条件WW的直接结果。后一条可证明如下：即$r\geqslant3$，r为整数。记$c_{ni}=b_{ni}/\sqrt{n}$，$c_n=\max\limits_{1\leqslant i\leqslant n}|c_{ni}|$，$d_{ni}=c_{ni}/c_n$，$i=1,\cdots,n$。则

$\sum\limits_{i=1}^n c_{ni}^2=1$（由$B_{2n}=n$推出）

$c_n\to0$（由$\{(b_{n1},\cdots,b_{nn}):n=1,\ 2,\ \cdots\}$满足条件N推出），

$|d_{ni}|\leqslant1$，$i=1,\ \cdots,\ n$，

现有

$$\left|\sum_{i=1}^n b_{ni}^r\right|\leqslant\sum_{i=1}^n|b_{ni}|^r=n^{r/2}\sum_{i=1}^n|c_{ni}|^r=n^{r/2}c_n^r\sum_{i=1}^n|d_{ni}|^r,$$

因$r>2$而$|d_{ni}|\leqslant1$，有$|d_{ni}|^r\leqslant d_{ni}^2$，故

$$\left|\sum_{i=1}^n b_{ni}^r\right|\leqslant n^{r/2}c_n^r\sum_{i=1}^n d_{ni}^2=n^{r/2}c_n^{r-2}\sum_{i=1}^n(c_nd_{ni})^2$$

$$=c_n^{r-2}n^{r/2}\sum_{i=1}^n c_{ni}^2=c_n^{r-2}n^{r/2}.$$

由于$c_n\to0$，证明了$B_{rn}=o(n^{r/2})$。

现在据(5A·2)有$l_n=0$，$\sigma_n^2=\dfrac{n^2}{n-1}$，而$(L_n-l_n)/\sigma_n=L_n\cdot\sqrt{\dfrac{n-1}{n}}$。把$L_n$写为(5·16)的形式，其中$a_i$要改为$a_{ni}$，而

$(\xi_1, \cdots, \xi_n)$以等概率 $1/n!$ 取$(b_{n1}, \cdots, b_{nn})$的任一置换。

有$E(L_n^r)=0$, 1, 当$r=1$, 2, 与(5A·1)的极限符合。取自然数$r\geqslant 3$. 有(记 $\mu_{nr}=E\{(L_n\sqrt{n-1}/n)^r\}$)

$$\frac{n^r}{(n-1)^{r/2}}\mu_{nr}=E(L_n^r)=E(a_{n1}\xi_1+\cdots+a_{nn}\xi_n)^r$$
$$=\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\cdots\sum_{i_r=1}^{n}a_{ni_1}\cdots a_{ni_r}E(\xi_{i_1}\cdots\xi_{i_r})$$
$$=\sum_{m=1}^{r}\sum_{e_1,\cdots,e_m}^{*}c(r;e_1,\cdots e_m)A_{e_1\cdots e_m}E(\xi_1^{e_1}\cdots\xi_m^{e_m}), \tag{5A·4}$$

此处

$$A_{e_1\cdots e_m}=\sum_{i_1,\cdots,i_m}' a_{i_1}^{e_1}\cdots a_{i_m}^{e_m}$$

(注意 $A_{e_1\cdots e_m}$ 与 n 有关, 省略了足标n)

Σ' 表示求和的范围为: i_1, …, i_m 是 1, …, n 中任意取 m个不同的数构成的全部排列。而

$C(r;e_1,\cdots,e_m)=\{$把 r 个相异物件分成 m 堆, 各堆物件个数为 e_1, …, e_m 且不计次序之不同分法(例如, $r=5$, $e_1=2, e_2=3$. 则{(A,B),(C,D,E)}与{(C,D,E),(A, B)}表示一种分法而非两种)},

$\sum\limits_{e_1,\cdots,e_m}^{*}$ 表示对所有这样的自然数$(e_1,\cdots, e_m)$求和: $e_1+\cdots+e_m=r$. 在得出(5A·4)时我们用到了根据$(\xi_1, \cdots, \xi_n)$的对称性(即对任一置换 i_1, …, i_n, $(\xi_{i_1},\cdots, \xi_{i_n})$与$(\xi_1, \cdots, \xi_n)$同分布)而得出的下述事实:

$E(\xi_{i_1}^{e_1}\cdots\xi_{i_m}^{e_m})=E(\xi_1^{e_1}\cdots\xi_m^{e_m})$, 当 $i_1,\cdots,i_m$ 两两不同. 记

$$B_{e_1\cdots e_m}=\sum_{e_1\cdots e_m}' b_{i_1}^{e_1}\cdots b_{i_m}^{e_m},$$

$\sum\limits_{e_1\cdots e_m}'$ 的意义与前同, 则

$$E(\xi_1^{e_1}\cdots\xi_m^{e_m})=\{n(n-1)\cdots(n-m+1)\}^{-1}B_{e_1\cdots e_m}$$
$$=(1+o(1))n^{-m}B_{e_1\cdots e_m},$$

此处 $o(1)\to 0$, 当 $n\to\infty$. 由上式及(5A·4), 得

$$\mu_{nr}=(1+o(1))\sum_{m=1}^{r}\sum_{e_1\cdots e_m}^{*} n^{-(m+\frac{r}{2})}A_{e_1\cdots e_m}B_{e_1\cdots e_m}$$
$$\cdot c(r;e_1,\cdots,e_m), \qquad (5A\cdot 5)$$

现欲证

$$\lim_{n\to\infty} n^{-(m+\frac{r}{2})}A_{e_1\cdots e_m}B_{e_1\cdots e_m}$$
$$=\begin{cases}1, & \text{若 } r \text{ 偶数}, m=r/2, e_1=\cdots=e_m=2\\ 0, & \text{其他情况}\end{cases} \qquad (5A\cdot 6)$$

若（$5A\cdot 6$）已证，则由之立即得出

$$\lim_{n\to\infty}\mu_{nr}=\begin{cases}C(r;\ 2,\ \cdots,\ 2), & \text{当 } r \text{ 为偶数；}\\ 0, & \text{当 } r \text{ 为奇数，}\end{cases} \qquad (5A\cdot 7)$$

而易见 $C(r;\ 2,\ \cdots,\ 2)=(r-1)!!$，当 r 为偶数。事实上，因堆不计次序，第一个物件所在的堆，可在剩下的 $r-1$ 个中任选一个与之配合，选法有 $r-1$ 种。故用归纳法立即得出上述结果。由此结果及（$5A\cdot 7$），立即得出（$5A\cdot 1$），而定理得证。

为证（$5A\cdot 6$），考察 $A_{e_1}\cdots_{e_m}$. 前面已记 $A_{rn}=\sum_{i=1}^{n}a_{ni}^{r}$. $A_{e_1}\cdots_{e_m}$ 可表为一些形如 $A_{j_1n}\cdots A_{j_hn}$ 的项的线性组合。此处 $j_1, \cdots, j_h$ 皆为自然数，和为 $j_1+\cdots+j_m=r$，而 $h\leqslant m$. 注意到 $A_{1n}=0$，可设 $j_1\geqslant 2, \cdots, j_h\geqslant 2$. 分两种情况：1. $m<r/2$. 这时由（$5A\cdot 3$）第一式，有 $A_{j_1n}\cdots A_{j_hn}=O(n^m)$. 2. $m\geqslant r/2$. 这时，利用 $j_1, \cdots, j_h$ 皆 $\geqslant 2$ 而和为 r，由（$5A\cdot 3$）第一式知 $A_{j_1n}\cdots A_{j_hn}=O(n^{r/2})$. 特别，若 $m=r/2$ 而 $j_1=\cdots=j_h=2$，则 $A_{j_1n}\cdots A_{j_hn}=n^{r/2}$. 对 $B_{e_1\cdots e_m}$ 如法炮制，令 $B_{rn}=\sum_{i=1}^{n}b_{ni}^{r}$. 把 $B_{e_1\cdots e_m}$ 表为一些形如 $B_{j_1n}\cdots B_{j_hn}$ 的项的线性组合，因为 $B_{1n}=0$，知可设 $j_1, \cdots, j_h$ 都 $\geqslant 2$. 仿上述推理，利用（$5A\cdot 3$）第二式，知 $B_{j_1n}\cdots B_{j_hn}=0(n^{r/2})$ 当 $m<r/2$（仍须利用 $h\leqslant m$）. 若 $m>r/2$，则由 $j_1,\cdots,j_h$ 皆 $\geqslant 2$ 知 $h<m$，故 $B_{j_1n}\cdots B_{j_hn}=o(n^m)$. 当 $m=r/2$ 时（这时 r 必为偶数），只有在 $h=m$ 而且 $j_1=\cdots=$

$j_h=2$ 时，才有 $A_{j_{1n}}\cdots A_{j_{nn}}=n^m$，否则 $A_{j_{1n}}\cdots A_{j_{hn}}=o(n^m)$。综合上述即得（$5A\cdot6$）．定理证毕。

二、定理 4·4 的一个较弱形式

利用定理 5·1 不难证明下面的结果：

定理 4·4′ 设把定理 4·4 的条件（1）强化为："$\{(C_{n1}, \cdots, C_{nn}):n=1, 2, \cdots\}$满足条件 WW"，则定理 4·4 的结论成立。

因为由条件 WW 可推出条件 N，故本定理比定理 4·4 弱一些，但在许多有关秩方法的应用中，本定理的条件常能满足。如在两样本问题中，有（$C_{n1}, \cdots, C_{nn}$）=（$0,\cdots,0,1, \cdots, 1$），其中 0 有 n_1 个，1 有 n_2 个，$n_1+n_2=1$．前已指出，若存在 $\lambda>0$ 使对一切 n 有 $\lambda\leqslant n_1/n\leqslant1-\lambda$，则条件 WW 满足。在一般两样本问题中，这条件总可认为是满足的．形式上说，为满足定理 4.4 的条件（1），只须 $n_1\to\infty$，$n_2\to\infty$．

为证明定理 4·4′，根据定理 5·1，只须证下述结论：若 $\varphi(\cdot)$ 为定义在（0，1）区间的非常数的平方可积计分函数，而 $C_{ni}=\varphi\left(\frac{i}{n+1}\right)$，$i=1, \cdots, n$，则 $\{(C_{n1},\cdots, C_{nn}):n=1, 2, \cdots\}$

满足条件 N．事实上，因为 $\varphi=\varphi_1-\varphi_2$，其中 φ_1 和 φ_2 都是非降的平方可积函数，故不失普遍性可设 φ 本身非降。记

$$\overline{\varphi}=\int_0^1\varphi(u)du$$

则由 φ^2 在（0，1）可积，易见

$$\frac{1}{n}\sum_{i=1}^{n}(C_{ni}-\overline{C}_n)^2\longrightarrow\int_0^1(\varphi(u)-\overline{\varphi})^2du>0, \quad (5A\cdot8)$$

上述积分大于 0 是根据 φ 在（0，1）不恒为常数的假定。又由 φ 非降，知

$$\max_{1\leqslant i\leqslant n}(C_{ni}-\overline{C}_n)^2\leqslant\left(\varphi\left(\frac{n}{n+1}\right)-\varphi\left(\frac{1}{n+1}\right)\right)^2$$

$$\leqslant2\varphi^2\left(\frac{n}{n+1}\right)+2\varphi^2\left(\frac{1}{n+1}\right), \quad (5A\cdot9)$$

有两种情况：一种是 φ 在 1 的附近有界。这时 $\varphi^2\left(\frac{n}{n+1}\right)=O(1)$ $=o(n)$．一种是 $\lim\limits_{u\to 1}\varphi(u)=\infty$.这时当$n$充分大时有 $\varphi\left(\frac{n}{n+1}\right)$ >0．故 $\int_{\frac{n}{n+1}}^{1}\varphi^2(u)du\geqslant\frac{1}{n+1}\varphi^2\left(\frac{n}{n+1}\right)$.但另一方面，由 φ^2 可积又有 $\lim\limits_{n\to\infty}\int_{\frac{n}{n+1}}^{1}\varphi^2(u)du\to 0$ 当 $n\to\infty$．这证明了 $\varphi^2\left(\frac{n}{n+1}\right)=$ $o(n)$．综合这两种情况都有 $\varphi^2\left(\frac{n}{n+1}\right)=o(n)$．同理证明 $\varphi^2\left(\frac{1}{n+1}\right)=o(n)$．由（$5A\cdot 9$），得

$$\max_{1\leqslant i\leqslant n}(C_{ni}-\overline{C}_n)^2=o(n), \qquad (5A\cdot 10)$$

把（$5A\cdot 8$）和（$5A\cdot 10$）结合，即得

$$\lim_{n\to\infty}\left\{\max_{1\leqslant i\leqslant n}(C_{ni}-\overline{C}_n)^2/\sum_{i=1}^{n}(C_{ni}-\overline{C}_n)^2\right\}=0,$$

于是$\{(C_{n1},\ \cdots,\ C_{nn}):n=1,\ 2,\ \cdots\}$满足条件 N．定理证毕。

三、（5·53）式的证明

这个证明用到概率论中之一周知的结果：若 ξ_1，ξ_2，…为一串 m维的随机向量，a 为 m 维常向量而 Λ 为 m 阶非负定方阵.若对任意的 m 维非 o 常向量 λ 都有 $\lambda'\xi_n\xrightarrow{\mathscr{L}}N(\lambda'a,\ \lambda'\Lambda\lambda)$，则必有 $\xi_n\xrightarrow{\mathscr{L}}N(a,\ \Lambda)$．

根据这个定理，为证（5·53），只需证明：对任何 $c-1$ 维非o 常向量 λ，有

$$\widetilde{S}_n^{-1}\lambda'\Lambda^{-1/2}\sum_{j=1}^{n}(\xi_j-E\xi_j)\xrightarrow{\mathscr{L}}N(0,\ \lambda'\lambda) \qquad (5A\cdot 11)$$

不妨设 $\lambda'\lambda=1$，无损于普遍性，

令

$$\eta_{nj}=\widetilde{S}_n^{-1}\lambda'\Lambda^{-1/2}(\xi_j-E\xi_j),\ j=1,\ \cdots,\ n,$$

则 η_{n1}，…，η_{nn} 为一串期望为 0 的独立随机变量，又

$$\mathrm{Var}(\eta_{nj})=\check{S}_n^{-2}\lambda'\Lambda^{-1/2}\Lambda S_j^2\Lambda^{-1/2}\lambda=\check{S}_n^{-2}S_j^2$$

此处用了 $\lambda'\lambda=1$（记号均见（5·53）式前面几行）。于是

$$B_n^2=\sum_{j=1}^n \mathrm{Var}(\eta_{nj})=\check{S}_n^{-2}\sum_{j=1}^n S_j^2=1,$$

又据假定，$\check{S}_n\to\infty$ 且 $\{\xi_1,\ \xi_2,\ \cdots\}$ 一致有界。故

$$\lim_{n\to\infty}\max_{1\leqslant j\leqslant n}|\eta_{nj}|=0, \tag{5A·12}$$

因此

$$\sum_{j=1}^n E|\eta_{nj}|^3\leqslant\max_{1\leqslant j\leqslant n}|\eta_{nj}|\sum_{j=1}^n E|\eta_{nj}|^2$$

$$=\max_{1\leqslant j\leqslant n}|\eta_{nj}|\sum_{j=1}^n \mathrm{Var}(\eta_{nj})=\max_{1\leqslant j\leqslant n}|\eta_{nj}|, \tag{5A·13}$$

由（5A·12）和（5A·13），有

$$\lim_{n\to\infty}\left\{\sum_{j=1}^n E|\eta_{nj}|^3/B_n^{3/2}\right\}=0,$$

于是，根据 ЛЯПУНОВ 中心极限定理，即知当 $n\to\infty$ 时

$$\sum_{j=1}^n \eta_{nj}\xrightarrow{\mathscr{L}}N(0,\ 1)$$

此即（5A·11），于是证明了所要结果。

习　题

5-1　根据 §5·1 中的几个例子，总结出：在何种情况下可以使用由全部样本作无限制置换而构成的置换检验。并以此说明："对称中心 $\theta=0$" 这个原假设不能用无限制的置换检验去检验之。

5-2　设 X_1，…，X_{n_1} 和 Y_1，…，Y_{n_2} 分别是抽自分布 $F(x)$ 和 $F(x/\theta)$ 的简单样本，分布 F 未知。试构造检验问题"$\theta=1\longleftrightarrow\theta>1$" 的一个置换检验。

5-3　考虑例 5·1 的一般情况，A、B 分别取 k 和 l 个水平。把 $k\times l$ 列联表写作如下形式：

B \ A	B_1	$\cdots$	B_j	$\cdots$	B_l	和
A_1	X_{11}	$\cdots$	X_{1j}	$\cdots$	X_{1l}	M_1
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
A_i	X_{i1}	$\cdots$	X_{ij}	$\cdots$	X_{il}	M_i
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
A_k	X_{k1}	$\cdots$	X_{kj}	$\cdots$	X_{kl}	M_k
和	N_1	$\cdots$	N_j	$\cdots$	N_l	n

其中 X_{ij} 是在 n 次独立观察中，A 取 A_i、B 取 B_j 的次数，$M_1, \cdots, M_k$ 和 $N_1, \cdots, N_l$ 分别是行和与列和。

证明　在原假设“A、B 两属性独立”成立时，在给定 $\xi = (M_1, \cdots, M_{k-1}, N_1, \cdots, N_{l-1})$ 的条件下，矩阵 $(X_{ij})_{i=1,\cdots,k-1,\, j=1,\cdots,l-1}$ 的条件分布只与 ξ 有关。这就是说，若以 p_i 和 q_j 分别记 A_i 和 B_j 的概率（当原假设成立时，(A_i, B_j) 的概率为 $p_i q_j$），则上述条件分布不依赖于 $p_1, \cdots, p_k$ 与 $q_1, \cdots, q_l$。

此结果可以像例 5·1 那样通过计算条件概率直接证明。另一个证法是用归纳法，即证明如题中之结论对 $k \leqslant k', l \leqslant l'$ 且 $k'+l' \leqslant k+l-1$ 时成立，则必对 k，l 成立。

5-4　举例说明：若在定理 5·1 中只假定两序列都满足条件 N，则（5·23）可以不成立。一个例子如下：取 $(a_{n1}, \cdots, a_{nn}) = (b_{n1}, \cdots, b_{nn}) = (1, \cdots, 1, 0, \cdots, 0)$，其中 1 有 $[n^{1/5}]$ 个。

5-5　但是，也存在这样的例子，其中两序列都只满足条件 N 而没有一个满足 WW，但（5·23）依旧成立。利用定理 4·4 可举出这样的例子：取

$(a_{n1}, \cdots, a_{nn}) = (1, \cdots, 1, 0, \cdots, 0)$，1 有 $[n^{1/2}]$ 个

$$(b_{n1},\cdots,b_{nn})=\left(\frac{n+1}{1}\right)^{1/3},\ \cdots,\ \left(\frac{n+1}{k}\right)^{1/3},\ \cdots,\ \left(\frac{n+1}{n}\right)^{1/3},$$

试就此两序列根据定理 4·4 证明定理 5·1 的（5·23）式，并证明上述两序列都只满足条件 N 而非条件 WW。

第六章 概率密度估计，非参数回归与判别

§6·1 概率密度估计

概率密度函数，常简称为密度函数以至密度，是概率论的最重要概念之一。虽然在统计学上我们常提“总体分布”这个名词，其实，使用密度的概念去规定或刻划一个统计模型不仅常见，且比使用分布概念更合适和方便。只要想想下面这个情况：在各种实际问题中，变量取值的分布呈现“两头小、中间大，左右对称”这种“正态类似型”者，为数颇多。这些特点在密度函数的图象上一目了然，而在分布函数的图象上则不然。

密度估计问题，就是要通过从总体中抽得的样本去估计其概率密度函数 f. 这里，估计的对象是一未知函数 f. 但在实际操作中，总可把问题说成：固定一已知的 x 值，要估计 f 在 x 点之值 $f(x)$。后者是一实数.于是，我们可以在习知的意义下，谈论密度的点估计、区间估计等等。

密度估计在统计上应用甚多，现在来看第二章中讨论过的用样本中位数$\hat{m}$去作总体中位数m的区间估计问题。若用 $\hat{m}$ 的渐近正态性并使用大样本区间估计，则需要作出密度 f 在 m 点的值 $f(m)$ 的估计.在这个及类似的例子中，需要作出未知密度 f 在一个点以至一定范围内取值的数值估计。在有些问题（特别是在有关选定统计模型的问题）中，只需要对密度图象的特征有所了解。例如，有一批观察数据，考虑用正态模型去分析之，则需要检验一下正态模型是否可用。有一些拟合优度检验可用于此目的。但一个在直观上更易理解和被接受的方法是：作出总体密度

估计的图象。若此图象大体上具备"两头低、中间高，左右对称"的特点，则对使用正态模型感到比较放心。当然，一般说来，密度估计的图象是通过对密度函数作数值估计，再用之作图得到的。下面是一个实际的例子。Bowyer 在 1980 年一项工作中，观察同型号的钢球高度，得到 15000 个数据。基此构造钢球高度的一个密度估计，其图象如图6·1。从图形可见：高度的分布是偏态的，有一个很长的低尾部。而分布的低尾部对应着钢球的凹陷处。这个估计显示出了正态分布之不适合。由于密度估计图象很直观和易于理解，从而可以解释数据，成为印证或支持某些科学技术结论的重要工具。

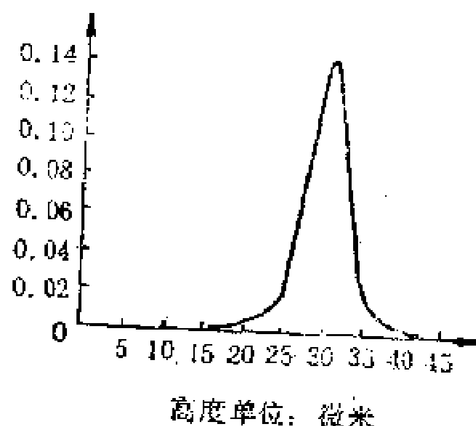


图6·1 钢球高度的密度估计

当谈到密度估计时，我们总是指未知密度函数 f 的所属类型并不知道的情况。当然，我们可以施加某些一般性的限制，如未知密度为连续的、单峰的，或在一定区间之外为 0 等等。因此，这是一个典型的非参数统计问题。这可以从反面去理解：设想我们已知或认定未知密度 f 属于正态类型，则 f 只取决于两个参数——期望 μ 和方差 σ^2。这时，与其去谈估计 f 的问题，不如说成是估计这两个参数的问题更简便。[1]

一、几种重要的密度估计方法

密度估计的方法很多。这里我们按照历史演变的顺序选择几种在应用上较重要的加以介绍。

1. 直方图法

---

① 若以 $f(x;\mu,\sigma^2)$ 记 $N(\mu,\sigma^2)$ 的密度，作出 μ 和 σ^2 的估计 $\hat{\mu}$ 与 $\hat{\sigma}^2$，可用 $f(x;\hat{\mu},\hat{\sigma}^2)$ 去估计 $f(x;\mu,\sigma^2)$，不过，我们也可以直接把 $f(x;\mu,\sigma^2)$ 作为估计对象。例如，$f(x;\mu,\sigma^2)$ 的最小方差无偏估计并非 $f(x;\overline{X},S^2)$，尽管 $\overline{X}$ 和 S^2 分别是 μ 和 σ^2 的最小方差无偏估计，这样做在实际问题中获益不多，故不大值得采用，因其估计方法大为复杂化了。无论如何，这并没有超出参数估计的范围。

九

此法基于概率密度的一个基本性质：随机变量 X 如有密度 f，则 X 取值在区间 $[a,b]$ 的概率 $P(a\leqslant X\leqslant b)$ 等于 $\int_a^b f(x)dx$. 若有 X 的简单样本 X_1, …, X_n，则 $P(a\leqslant x\leqslant b)$ 可用

$$\#\{i:1\leqslant i\leqslant n,\ a\leqslant X_i\leqslant b\}/n$$

去估计。因此，$P(a\leqslant X\leqslant b)/(b-a)$ 即

$$\int_a^b f(x)dx/(b-a) \text{ 可以用}$$

$$\#\{i:1\leqslant i\leqslant n,\ a\leqslant X_i\leqslant b\}/n(b-a)$$

去估计：当 $b-a$ 充分小时，$\int_a^b f(x)dx/(b-a)$ 可近似代表 $f(x)$ 在区间 $[a,b]$ 上之值。这样就得到了 f 的一个估计。基于上述原理，这方法可描述如下：选择一个适当的正数 h. 把全直线分为一些长为 h 的区间。任取这些区间之一，记为 I. 对 $x\in I$，以

$$\#\{i:1\leqslant i\leqslant n,\ X_i\in I\}/nh \tag{6·1}$$

作为 $f(x)$ 的估计。这个估计的图形是一个边长为 h 的阶梯形。若从每一端点向底边作垂线以构成矩形，则得到一个如图 6·2 的图形。它是由一些直立的矩形排在一起而成的，以此得到直方图之名。

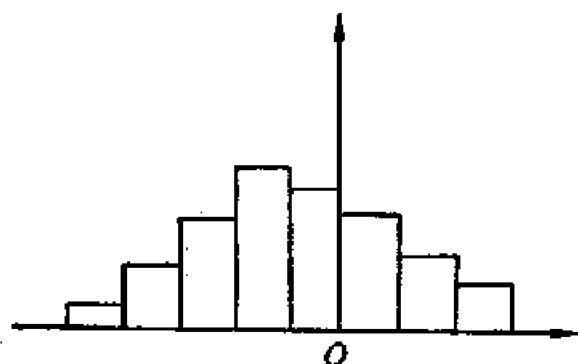


图6·2　直方图

在图 6·2 中，O 是分割点之一。当然也可预先选定任一点 a 为分割点。这时所有分割点都有 $a+ih$ 的形式，$i=0$，±1，±2，…. 重要的是 h 的选择。h 太大了，平均化的作用突出了，而淹没了密度的细节部分。太小了，则受随机性影响太大，而产生极不规则的形状。h 的选择无现成规则可循。一般只能说，应选择一个适当的 h 以平衡上述两种效应。总的讲，当样本大小 n 大时，h 可取得小一些。

在关于直方图的理论讨论中，我们常假定区间分割（即上文 a、h 的选择）是在考察样本之前就定下来的，因此无随机性。这就使理论简化了。但在实际操作时不一定能恪守这个规定。例如，一批样本可能较集中在 O 点附近，而在较远的地方个数较少。这时，有条件把密度 f 在 O 点附近之值估计得细一些，而在远处则只能满足于较粗的估计。就是说，我们可能取一些不等长的区间，区间长度在 O 附近很短而在远离 O 点处则较长。然后在每一区间工内按（6·1）式作出 f 的估计。这时，区间的位置、长短都是在参考了样本以后决定的，故有随机性。这样的直方图估计称为"Data- based"的直方图估计。其理论较 a 和 h 都比随机的通常直方图估计复杂得多，本书将不加讨论。

直方图估计的优点在于简单易行，且在 n 较大因而容许 h 较小的情况下，所得图象尚能显示密度的基本特征。但也有明显的缺点。它不是连续函数(这可以通过适当地修匀来解决)，且从统计角度看一般说效率较低。例如，在这一方法下，每一区间中心部分密度估计较准，而边缘部分则较差。综合种种因素，我们仍可以说：直方图估计不失为一个有用而基本的密度估计方法。

2．Rosenblatt 法

为克服上文提到的直方图法的一个缺点——对每个区间边缘部分密度值的估计较差，Rosenblatt 在 1955 年提出了一个简单的改进。指定一个正数 h 如前，对每个 x，以 I_x 记以 x 为中心，长为 h 的区间，即 $\left[x-\frac{h}{2},\ x+\frac{h}{2}\right]$。以 I_x 作为（6·1）式中之 I，算出之值作为 f 在 x 点处之值 $f(x)$ 的估计。这就是 Rosenblatt 估计。我们用 $f_n(x) \triangleq f_n(x;\ X_1,\ \cdots,\ X_n)$ 表示这个估计，则有

$$f_n(x)=\frac{1}{nh}\#\{i:1\leqslant i\leqslant n,\ X_i\in I_x\}. \qquad (6\cdot2)$$

Rosenblatt 法与直方图法不同之处仅在于，它事先不把分割区间定下来，而让区间随着要估计之点 x 跑，使 x 始终处在区间之

中心位置，而获致较好的效果。理论上可以证明：从估计量与被估计量接近的数量级上看，Rosenblatt 方法确实优于直方图法。

3．Parzen 的核估计

细心的读者不难看出：Rosenblatt 估计仍为一阶梯函数，只不过与直方图估计比起来，各阶梯之长不一定相同而已，仍非连续曲线。另外，从 Rosenblatt 估计的定义中看出，为估计 f 在 x 点之值 $f(x)$，对与 x 在一定距离（确切地说，是 $h/2$）内的样本，起的作用一样，而在此以外则毫不起作用。直观上可以设想：为估计 $f(x)$，与 x 靠近的样本，所起作用似应比远离 x 的样本要大些。这些在 Parzen 于 1962 年提出的核估计方法中都得到了体现。

为介绍 Parzen 的思想，我们先将（6·2）式变换一个形式，引进一个函数

$$W(x)=I_{[-\frac{1}{2},\frac{1}{2}]}(x)=\begin{cases}1, & 当 -\frac{1}{2}\leqslant x<\frac{1}{2}\\ 0, & 其它\ x,\end{cases} \tag{6·3}$$

则（6·2）式可改写为

$$f_n(x)=\frac{1}{nh_n}\sum_{i=1}^{n}W\left(\frac{x-X_i}{h_n}\right). \tag{6·4}$$

（6·3）定义的 W 是 R^1 上的密度函数，但是一种特殊的密度函数，即均匀密度。Parzen 的推广即在于去掉这一特殊性，而容许 W 可以为一般的密度函数。下面我们给 Parzen 的核估计下一个正式的定义。

定义6·1 设 $K(\cdot)$ 为 R^1 上的一个给定的概率密度函数，$h_n>0$ 是一个同 n 有关的常数，定义

$$f_n(x)=\frac{1}{nh_n}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h_n}\right), \tag{6·5}$$

称 f_n 为总体未知密度 f 的一个核估计，K 为核函数，h_n 为窗宽。

这一定义考虑的是 X 为一维的情况。若 X 为 d 维，只须

将（6·5）式中分母 nh_n 改为 nh_n^d. 又从定义可见，Rosenblatt 估计（6·2）是核估计的一个特例，其中 $K(\cdot)$ 由（6·3）式所确定。这里需要对上述定义作几点注解.

（1）"窗宽"（Window-width）这个词是从核估计的特殊形式（6·3）、（6·4）中 h_n 的含义派生出来的。我们从公式(6·3)（6·4)可以解释 Rosenblatt 估计为：对每个观察 X_i 限制在高为 $\frac{1}{nh_n}$，宽为 h_n 的"窗"内，而估计值为 n 个这种"窗"之和。因而 h_n 正是这 n 个"窗"的公共"窗宽"参数。

（2）窗宽 h_n 的作用。由定义可知，核估计既同样本有关，又同核 K 及窗宽 h_n 的选取有关。在给定样本之后，一个核估计性能的好坏，取决于核及窗宽的选取是否适当。从直观上看，核估计在每观察点 X_i 有一"碰撞"，估计量是这些"碰撞"之和，核 K 确定了每一个"碰撞"的形状，而 h_n 则决定了"碰撞"的宽度，当 h_n 选得过大，由于 x 经过压缩变换 $\frac{x-X_i}{h_n}$ 之后使分布的主要部分的某些特征（如多峰性）被掩盖起来了，估计量有较大偏差；如 h_n 太小，整个估计特别是尾部出现较大的干扰，从而有增大方差的趋势。因而在实际使用核估计时，如何选取适当的宽度是一项很细致的工作。

（3）从理论上讲，关于核 K 的要求尚可适当放宽。即不一定要求 K 为密度，甚至也不必要求它为非负。但从实用上看，要求 K 为概率密度函数是合适的。这是因为待估的 f 是密度，最好是估计量 f_n 本身也是密度函数。当 K 为密度时，容易验证 f_n 满足这个条件。而且当 K 满足某些光滑条件时，f_n 作为 x 的函数，同样继承这些光滑性质。从而可以弥补 Rosenblatt 估计的不足，选择核 K 是否适当，同样要影响估计的精度。原则上，我们可对核 K 施加一定的限制，使得估计量与待估函数的偏差在一定意义下尽可能地小。例如可以要求 K 有对称性，其一阶矩（关于密

度K）为零，具有有界性、连续性等等。

往后将会看到核估计有种种优良性质，且便于理论分析。因而在文献中，核估计已成为密度估计的主要方法。

4．最近邻估计

在文献中，除核估计外，最近邻估计方法也是常用的一种密度估计方法。这是 Loftsgarden 和 Quesenberry 在 1965 年提出的。此法较适合于密度的局部估计。其要旨如下：设 X_1，…，X_n是来自未知密度 f 的样本。先选定一个同 n 有关的整数 $k=k_n$，合于$1\leqslant k<n$，对固定的 $x\in\mathbf{R}'$，记 $a_n(x)$为最小的正数 a 使得$[x-a,\ x+a]$中至少包含 X_1，…，X_n中的 k个。注意到，对每一 $a>0$ 可以期望在 X_1，…，X_n中大约有 $2anf(x)$ 个观察值落入区间$[x-a,\ x+a]$之中。因而值 $f(x)$的估计（记为$\hat{f}_n(x)$）自然地可以通过令 $k=2a_n(x)n\hat{f}_n(x)$得到。于是定义

$$\hat{f}_n(x)=k_n/2a_n(x)n \tag{6·6}$$

为 $f(x)$的估计。文献上称$\hat{f}_n$为 f 的最近邻估计（简记为N.N．估计）。注意到与 Rosenblatt 估计相反，此处区间长度$2a_n(x)$是随机的，而区间内所含观察数是固定的。下面的引理说明：从整体上看，N. N. 估计的性质与核估计有很大的不同。

引理6·1 （1）对固定 n 及 X_1，…，X_n，$\hat{f}_n(x)$作为变元 x 的函数是处处连续的。

（2）
$$\int\hat{f}_n(x)dx=\infty.$$

证 （1）任取 $x\in\mathbf{R}^1$，$x'\in\mathbf{R}^1$，不失一般性可设 $x'\leqslant x$．则由 $a_n(\cdot)$的定义易知

$$a_n(x')\leqslant a_n(x)+(x-x')=a_n(x)+|x-x'|,$$
$$a_n(x')\geqslant a_n(x)-(x-x')=a_n(x)-|x-x'|。$$

因而

$$|a_n(x')-a_n(x)|\leqslant|x'-x|, \tag{6·7}$$

自得证第一个结论。

(2) 记 $X_{(1)}<X_{(2)}<\cdots<X_{(n)}$ 为 $X_1, \cdots, X_n$ 的次序统计量。当 $x<X_{(1)}$ 时，$a_n(x)=X_{(k)}-x$；当 $x>X_{(n)}$ 时，$a_n(x)=x-X_{(n-k+1)}$. 因而

$$\int \hat{f}_n(x)\,dx \geqslant \left[\int [x<X_{(1)}]+\int [x>X_{(n)}]\right]\hat{f}_n(x)\,dx$$

$$=\int_{-\infty}^{X_{(1)}} \frac{k}{2n\,(X_{(k)}-x)}\,dx$$

$$+\int_{X_{(n)}}^{\infty} \frac{k}{2n\,(x-X_{(n-k+1)})}\,dx=\infty.$$

引理证毕。

由引理 6·1 可知：$\hat{f}_n(x)$作为变元 x 的函数非概率密度。另外，从证明过程可看出：

$$\hat{f}_n(x)=O\left(\frac{1}{n}\right),\ \text{当}\ |x|\to\infty.$$

注意到这一性质与待估 f 的尾部特征无关，因而对相当一类待估密度，估计 $\hat{f}_n(x)$ 的尾部衰减得太慢。从而 $\hat{f}_n$ 不适宜用作 f 的整体估计。下面的引理给出了 $\hat{f}_n(x)$ 的分布。大体上说来，Rosenblatt 估计与 N.N. 估计的关系犹如二项分布与负二项分布的关系，因而 N.N. 估计的性质显得复杂些。

引理6·2 对固定 $x\in R$，$n\geqslant 1$，有

$$P(a_n(x)\leqslant y)=\sum_{i=k}^{n}\binom{n}{i}p^i(y)(1-p(y))^{n-i} \qquad (6\cdot 8)$$

$$=n\binom{n-1}{k-1}\int_0^{p(y)} t^{k-1}(1-t)^{n-k}dt,$$

$$y>0, \qquad (6\cdot 9)$$

其中

$$P(y)=\int_{x-y}^{x+y} f(t)\,dt=P(x-y\leqslant X\leqslant x+y). \qquad (6\cdot 10)$$

证明留作练习。

由(6·9)式，即得 $a_n(x)$ 有概率密度

$$g_n(y) = \begin{cases} n\binom{n-1}{k-1} p(y)^{k-1}(1-p(y))^{n-k}[f(x+y)-f(x-y)], & y>0 \\ 0, & y\leqslant 0 \end{cases} \tag{6·11}$$

如果令

$$K(x)=\begin{cases} \frac{1}{2}, & |x|<1 \\ 0, & |x|\geqslant 1 \end{cases} \tag{6·12}$$

则可将（6·6）改写为

$$\hat{f}_n(x)=\frac{1}{na_n(x)}\sum_{i=1}^{n} K\left(\frac{x-X_i}{a_n(x)}\right). \tag{6·13}$$

因而对固定的 x，N.N. 估计可看成以（6·12）为核，具窗宽 $a_n(x)$ 的核估计。也就是说，在单个点 x 上的 N.N. 估计与核估计差之不大，只有当同时考虑在几个点或者估计整个 f 时，这两种方法才显示出差别。注意到 $a_n(\cdot)$ 在每一形如 $\frac{1}{2}[X_{(i)}+X_{(i+k)}]$（$1\leqslant i\leqslant n-k$）的点上其导数有间断，因而有局部干扰。而对核估计来说，只要有适当光滑的核，就可得到有相同光滑程度的核估计。但这里并不企图对这两种方法作全面的比较，因为这只有在进行深入的理论分析之后才能作出。N.N. 估计由于计算上有某种方便之处，这种方法被广泛地用于模式识别及非参数判别分析。在文献中，也有将 N.N. 估计与核估计结合起来成为（6·13）的一种推广形式，即（6·13）中的 K 为任一核函数，而不必有（6·12）的形式。这种推广的好处在于可通过适当选择核而改进估计量在尾部的性能。

二、估计精度的度量

我们用 $T_n(x)\triangleq T_n(x; X_1, \cdots, X_n)$ 表示基于样本 $X_1, \cdots, X_n$，未知密度 $f(x)$ 的任一估计。由于 $T_n(x)$ 既同样本有关，又是考察点的函数。因而对固定的考察点 x，估计精度的一种自然测度为

$$\mathrm{MSE}(T_n(x)) = E_f\{T_n(x) - f(x)\}^2, \quad (6\cdot14)$$

称（6·14）为估计 T_n 的均方误差，其中 E_f 表示期望是在真分布为 f 时计算。而当真分布较明确时，也简记 E_f 为 E。我们熟知有

$$\mathrm{MSE}(T_n(x)) = \{E_f(T_n(x)) - f(x)\}^2 + \mathrm{Var}_f(T_n(x)). \quad (6\cdot15)$$

上式右端是由两个部分组成：第一项是偏差项，而第二项是估计的方差。我们自然希望这两部分越小越好。但是要同时减少这两部分是困难的。通常，如降低偏差，则方差有增大的趋向，反之亦然。直观上看，偏差项表明估计量对 f 的光滑修正的程度。一个估计量的光滑程度越高，可能更多地忽略 f 的某些细节，从而增大随机误差。对于 $T_n(x)$ 为核估计时，有

$$E_f[T_n(x)] = \int K(y) f(x - h_n y)\, dy \quad (6\cdot16)$$

$$\mathrm{Var}_f[T_n(x)] = \frac{1}{nh_n} \int K^2(y) f(x - h_n y)\, dy - \frac{1}{n}\left\{\int K(y) f(x - h_n y)\, dy\right\}^2. \quad (6\cdot17)$$

因而一个核估计的光滑程度只与光滑参数 h_n 有关（当核 K 已确定时），而与 n 无直接关系。为了降低其均方误差，必须调整光滑参数。

对于密度估计来说，更有实际意义的精度的度量应是整体性的测度。首先由 Rosenblatt（1956年）提出而后被广泛使用的一个整体测度是积分均方误差（MISE）：

$$\mathrm{MISE}(T_n) = E\left\{\int (T_n(x) - f(x))^2 dx\right\}, \quad (6\cdot18)$$

易知

$$\mathrm{MISE}(T_n) = \int E[T_n(x) - f(x)]^2 dx \quad (6\cdot19)$$

$$= \int \mathrm{MSE}[T_n(x)]\, dx$$

$$=\int[ET_n(x)-f(x)]^2dx$$

$$+\int \mathrm{Var}(T_n(x))\,dx. \qquad (6\cdot20)$$

因而

MISE＝积分偏差平方和＋积分方差。

由公式（6·20），我们在前段对均方误差的分析，同样可施用于积分均方误差。对于核估计来说，应该选择 h_n 使得相应的核估计其 MISE 达到最小。文献上称这种 h_n 为核估计的最佳窗宽。在实际问题中，如何选择最佳窗宽是个难以处理的问题。下面举一个例子，设K为标准正态密度，而 f 为 $N(\mu,\sigma^2)$ 密度。由（6·16）、（6·17）及（6·20）易得

$$E_f[T_n(\cdot)]=N(\mu,\sigma^2+h_n^2)\text{密度}$$

$$(2\sqrt{\pi})\ \mathrm{MISE}(T_n)=\frac{1}{n}[h_n^{-1}-(\sigma^2+h_n^2)^{-1/2}]$$

$$+\sigma^{-1}+(\sigma^2+h_n^2)^{-\frac{1}{2}}$$

$$-2\sqrt{2}(2\sigma^2+h_n^2)^{-\frac{1}{2}}. \qquad (6\cdot21)$$

再对（6·21）关于 h_n 求极小，即得最佳窗宽。Deheuvels 曾给出数值计算的结果，即使对于 $n=10$ 这样的小样本，在最佳窗宽下算得的 MISE 非常小。

为便于计算及理论分析，下面我们分别导出估计偏差及方差的渐近表达式。为简单计，设K是对称密度函数，满足：

$$\int tK(t)\,dt=0,\ k_2\triangleq\int t^2K(t)\,dt\neq0. \qquad (6\cdot22)$$

而 f 具有二阶有界连续导数，$h\triangleq h_n\to0$，当 $n\to\infty$。由公式（6·16），使用 $f(x-hy)$ 在 x 处的 Taylor 展开，有

$$E_f[T_n(x)]-f(x)=\int K(u)[f(x-hu)$$

$$-huf'(x)-f(x)]du$$

$$=h^2\int K(u)u^2 f''(x-\theta hu)du/2.$$

其中$|\theta|\leqslant 1$（θ 与 x、u、n 有关)。由对 f 的假设，使用控制收敛定理可得

$$E_f T_n(x)-f(x)=\frac{1}{2}f''(x)k_2h^2+o(h^2). \quad (6\cdot 23)$$

用同样的方法，可得

$$\mathrm{Var}_f(T_n(x))=\frac{1}{nh}f(x)\int K^2(u)du+o((nh)^{-1}). \quad (6\cdot 24)$$

因此当 f 满足上述条件，且 $f''\in L_2(\mathbf{R}^1)$时，有如下渐近公式：

$$\int\left[E_f T_n(x)-f(x)\right]^2 dx\approx\frac{1}{4}h^4k_2^2\int[f''(x)]^2dx, \quad (6\cdot 25)$$

$$\int \mathrm{Var}_f[T_n(x)]dx\approx(nh)^{-1}\int K^2(u)du. \quad (6\cdot 26)$$

从公式(6·25)、(6·26)可见：如 h 选得很小，固然可降低偏差，但方差项随之增大；反之亦然。今合并(6·25)、(6·26)，得到 MISE 的渐近公式：

$$\mathrm{MISE}\approx\frac{1}{4}h^4k_2^2\int[f''(x)]^2dx+(nh)^{-1}\int K^2(u)du. \quad (6\cdot 27)$$

再对(6·27)右端关于 h 求极小，得到渐近最佳窗宽（记为 h_{opt})：

$$h_{\mathrm{opt}}=k_2^{-2/5}\left[\int K^2(u)du\right]^{1/5}\left[\int(f''(x))^2dx\right]^{-1/5}n^{-1/5}. \quad (6\cdot 28)$$

公式(6·28)表明：最佳渐近窗宽随 n 增大以 $n^{-1/5}$ 的速度趋于零。其次，积分$\int[f''(x)]^2dx$ 直观上可看成是 f 的振动频率的一种度量。因而对于摆动频率大的 f，其最佳的 h 应该随之变小。但是，由于公式(6·28)含有未知的密度 f，尚不能付诸应

用。一种替代办法是估计积分$\int[f''(x)]^2dx$，再将估计量代入公式（6·28）。这样得到的窗宽已是样本的函数。另外一种方法就是直接由样本"自动"选择窗宽，文献上称之为"自适型"核估计。但不论使用哪一种方法，得到的窗宽已失去"最佳窗宽"的原意了。

如将由（6·28）确定的h_{opt}代入（6·27），则有

$$\mathrm{MISE}\approx\frac{5}{4}C(K)\left\{\int[f''(x)]^2dx\right\}^{1/5}n^{-4/5} \qquad (6\cdot29)$$

其中

$$C(K)=k_2^{2/5}\left\{\int K^2(t)dt\right\}^{4/5}. \qquad (6\cdot30)$$

然后可依使$C(K)$尽可能小的原则（当然要满足（6·22））选择K。这样可以得到尽可能小的积分均方误差。至于满足这种要求的核的选择问题，文献上已有一些讨论，还有待于进一步发展。从公式（6·29）至少可看出这样一个事实：不论h及K如何选取，作为核估计来说，其积分均方误差收敛于零的速度，其主要部分的阶不能超过4/5。这在理论分析上是很有意义的。

三、密度估计的应用

密度估计是具有广泛应用领域的一种非参数统计方法。Silverman曾指出，密度估计在数据的统计处理的所有阶段都是有用的。其应用领域涉及社会科学、物理科学、生物科学以及各种工程技术领域。这里应指出的是，密度估计的重要性，并不在于它的单独使用，而是作为统计推断的中间环节发挥作用。下面就三个方面作一简单介绍。

1. 非参数判别

判别分析的基本问题可简单地表示为：设有来自总体A的样本X_1，…，X_n，及来自总体B的样本Y_1，…，Y_m。今有新的观察Z，问Z来自A还是B？现设总体A有密度f_A，B有密度f_B。基于极大似然原理可定出如下的判别规则：如果

$$f_A(Z) \geqslant f_B(Z)$$

则判 Z 属于总体 A，反之则判为 B。但在实际问题中，f_A 及 f_B 往往是未知的，这样的判别规则无实用价值。Fix 和 Hodges（1951）提出了一种非参数方法，即：分别基于 X_1，…，X_n 及 Y_1，…，Y_m 估计 f_A 及 f_B，记估计为 $\hat{f}_A$ 及 $\hat{f}_B$。然后视 $\hat{f}_A(Z) \geqslant \hat{f}_B(Z)$ 抑或 $\hat{f}_A(Z) < \hat{f}_B(Z)$ 确定 Z 所归属的类。这是一种最简单的非参数判别方法。当然还有别的非参数判别规则，而且非参数判别并非必须使用密度估计。这在后文（§6·4）中将要详细介绍。

2．聚类分析

设有 n 个来自未知密度 f 的观察 X_1，…，X_n，要求依某种规则将 X_1，…，X_n 分成若干类。与判别分析不同的是：关于类及类的数目不是事先给定的，而是要由这组观察来确定。在考古学中就有这样的问题。一种常用的聚类方法即是构造某种“树图”。各个个体（即 X_i）按“树图”中的等级归并成若干类，而划分等级的规则需使用密度估计。

3．随机数的模拟

设已有观察 X_1，…，X_n，由于随机影响，这些观察渗杂了某些伪造的细节。我们的目的是模拟一组新数据 Y_1，Y_2，…，使得 Y_1，Y_2，… 具有原总体的结构，但无这些伪造的细节。当总体具未知密度 f 时，可用其核估计产生模拟数，例如 $\hat{f}$ 是基于 X_1，…，X_n 的具核 K 及窗宽 h_n 的密度估计，可按以下步骤产生新数据 Y：

（1）从数字 1，2，…，n 中有放回地随机抽取一个，记为 I；

（2）产生一个与 $X_1,\cdots,X_n$ 独立的具密度 K 的随机变量 ε；

（3）定义

$$Y = X_I + h\varepsilon.$$

以上过程可不断地重复进行，从而产生一串新数据。易知这样的

Y 有分布密度 $\tilde{f}$.

其它的应用，例如多峰性检验，各种密度泛函估计等等，在此不一一列举。

§6·2 密度估计的大样本性质

在上一节中，我们讨论了估计概率密度的几种常用方法，并指出了其若干初步的小样本性质。由于对未知密度的数学形式没有任何假定，指望得出较为深入的小样本性质是不现实的。例如，这些估计都没有无偏性，也不知道它们是否在任何有意义的小样本优良性准则之下具有最优性。

由于这个原因，迄今为止关于密度估计的研究，几乎全集中在大样本方面。一般来说这本是非参数方法的一个特征。

一、有关概念

为下文的讨论方便起见，先回顾并叙述若干有关概念。以下总假定 X_1，…，X_n 是来自未知密度 f 的独立同分布样本，$T_n(x)=T_n(x; X_1, \cdots, X_n)$ 是基于该样本的 $f(x)$ 的任一估计。首先，类似于参数估计中的渐近无偏性，我们有

定义6·2 如果对每一给定 x

$$\lim_{n\to\infty} E_f(T_n(x))=f(x)，对所有可能的 f \qquad (6·31)$$

则称 T_n 为渐近无偏估计。

在文献中，已经证明了：在相当宽泛的条件下，对固定 n，密度函数的无偏估计是不存在的。直观上看，只要固定 n，由样本 X_1，…，X_n 所提供的关于 f 的信息总是有限的，即使估计方法不断变更也于事无补。但当 n 无限增大时，我们对 f 的了解也就逐渐完整。正如往后的讨论所表明的那样，在不太强的限制下，渐近无偏估计总是存在的。

其次，在参数估计中的相合性概念也可几乎是平行地移到这里。

定义6·3 如果对固定 x，有

$$\lim_{n\to\infty} E[T_n(x)-f(x)]^2=0 \tag{6·32}$$

则称 T_n 为 f 的（在 x 处）均方相合估计．简记为

$$T_n(x)\Rightarrow f(x).$$

显然，为证 $T_n(x)\Rightarrow f(x)$，只须证：当 $n\to\infty$时，

$$ET_n(x)\longrightarrow f(x) \text{ 及 } \mathrm{Var}_f(T_n(x))\longrightarrow 0.$$

类似可定义对固定 x，$T_n(x)$ 依概率收敛于 $f(x)$（记为 $T_n(x)\xrightarrow{p} f(x)$）及以概率1收敛（记为 $T_n(x)\longrightarrow f(x)$，a.s.）．以上仅就固定考察点 x，讨论估计量 $T_n(x)$向 $f(x)$ 的逼近，文献上称之为逐点相合性。与此相关的概念，则是一致相合性．现分别给出如下的定义。

定义6·4 如对任给的 $\varepsilon>0$

$$\lim_{n\to\infty} P(\sup_x |T_n(x)-f(x)|\geqslant\varepsilon)=0 \tag{6·33}$$

则称 T_n 是 f 的一致相合估计，并简记为

$$\sup_x |T_n(x)-f(x)|\xrightarrow{p} 0, \text{ 当} n\to\infty.$$

定义6·5 如果

$$P(\lim_{n\to\infty}\sup_x |T_n(x)-f(x)|=0)=1 \tag{6·34}$$

则称 T_n 为 f 的一致强相合估计，并简记为

$$\sup_x |T_n(x)-f(x)|\longrightarrow 0, \text{ a.s.，当} n\to\infty.$$

在定义6·4、6·5中，暗含着 $\sup_x |T_n(x)-f(x)|$ 作为样本 $X_1,\cdots,X_n$ 的函数是可测的．这对于几种常用的密度估计都是满足的。显然对一致相合性的要求要比逐点相合性高得多。通常证明(6·33)或（6·34）是分两步进行的．其一，是证明

$$\lim_{n\to\infty}\sup_x |ET_n(x)-f(x)|=0; \tag{6·35}$$

其二，是断定

$$\text{当 } n\to\infty \text{ 时}\quad \sup_x |T_n(x)-ET_n(x)|\xrightarrow[\text{(或a.s.)}]{p} 0. \tag{6·36}$$

这第一部分无随机性可言，完全由 f 及估计量的光滑性所确定，因而较容易。主要困难在第二部分。在某些情况下，可将 $\sup_x |T_n(x)-ET_n(x)|$ 表成经验过程的适当泛函，然后使用经验过程的有关性质得以证明。

二、核估计的大样本性质

本段总用(除非另有说明) K 表示 $\mathbf{R}^1$ 上核函数，h_n 为窗宽，$f_n(x)$ 为具有核 K、窗宽 h_n 的基于 $X_1, \cdots, X_n$ 的核估计，其定义同 (6·5)。我们讨论核估计的最基本且较为初等的若干大样本性质。下面的引理可以说是核估计的一个基本引理，最先是由 Parzen (1962) 给出的。

引理6·3 设 $K(\cdot)$ 及 $g(\cdot)$ 均为 $\mathbf{R}^1$ 上的 Borel 可测函数，满足下述条件：

(1) K 有界；

(2) $\int |K(u)|du<\infty$；

(3) $\lim_{|u|\to\infty} uK(u)=0$ 或 g 有界；

(4) $\int |g(u)|du<\infty$，

常数序列 $\{h_n\}$ 满足 $\lim_{n\to\infty} h_n=0$。令

$$g_n(x)=\frac{1}{h_n}\int K\left(\frac{u}{h_n}\right)g(x-u)du, \tag{6·37}$$

则

$$\lim_{n\to\infty} g_n(x)=g(x)\int K(u)du,\quad 任意\ x\in c(g). \tag{6·38}$$

其中 $C(g)$ 为 g 的连续点集。

又若 g 有界且一致连续，则 (6·38) 关于 x 一致成立

证 先设 $\lim_{|u|\to\infty} uK(u)=0$。取定 $\delta>0$，有

$$|g_n(x)-g(x)\int K(u)du|$$

$$=\left|\int[g(x-u)-g(x)]\frac{1}{h_n}K\left(\frac{u}{h_n}\right)du\right|$$

$$\leqslant\sup_{|u|<\delta}|g(x-u)-g(x)|\int|K(u)|du$$

$$+\sup_{|u|\geqslant\delta}\left|\frac{u}{h_n}K\left(\frac{u}{h_n}\right)\right|\int|g(u)|du$$

$$+|g(x)|\int_{|u|\geqslant\delta}\left|\frac{1}{h_n}K\left(\frac{u}{h_n}\right)\right|du$$

$$\triangleq J_{n1}+J_{n2}+J_{n3}. \qquad (6\cdot39)$$

因 $x\in C(g)$，对任给 $\varepsilon>0$，可选 $\delta>0$ 充分小使得 $J_{n1}<\varepsilon$，然后固定此 δ. 由 $\lim\limits_{n\to\infty}h_n=0$ 可知 $\lim\limits_{n\to\infty}J_{n2}=0$. 又

$$\int_{|u|\geqslant\delta}\left|\frac{1}{h_n}K\left(\frac{u}{h_n}\right)\right|du=\int_{|u|\geqslant\frac{\delta}{h_n}}|K(u)|du.$$

由条件（2）即得 $\lim\limits_{n\to\infty}J_{n3}=0$. 因而

$$\overline{\lim_{n\to\infty}}\sup_x\left|g_n(x)-g(x)\int K(u)du\right|\leqslant\varepsilon.$$

由 $\varepsilon>0$ 的任意性即得证（6·38）.

若 g 有界，记 $M=\sup\limits_u|g(u)|$. 则 $J_{n2}+J_{n3}$ 代之以

$$2M\int_{|u|\geqslant\delta}\left|\frac{1}{h_n}K\left(\frac{u}{h_n}\right)\right|du=2M\int_{|u|\geqslant\delta/h_n}|K(u)|du.$$

仍由条件（2）得到（6·38）. 至于第二个结论，只要注意到，由 g 一致连续，对任给 $\varepsilon>0$，可找到 $\delta>0$ 使（6·39）右端第一项关于 x 一致地小于 ε. 其余相同，引理证毕。

下面讨论核估计的逐点相合性。

定理6·1 设核 K 是 R' 上的概率密度，且满足引理 6·3 之条件（1）、（2）。若 $\lim\limits_{n\to\infty}h_n=0$，则有

$$\lim_{n\to\infty}Ef_n(x)=f(x),\quad x\in c(f) \qquad (6\cdot40)$$

又若 f 一致连续，则（6·40）关于 x 一致成立。

证 第一个结论是引理 6·3 的直接结果；又若 f 一致连续，则 f 有界。因而第二个结论立即可得，定理证毕。

定理6·2 设核K满足定理6·1的条件，且

$$\lim_{n\to\infty} h_n=0,\quad \lim_{n\to\infty} nh_n=\infty \tag{6·41}$$

则

$$f_n(x)\Rightarrow f(x),\quad x\in c(f). \tag{6·42}$$

证 固定$x\in c(f)$. 由定理6·1只须证

$$\mathrm{Var}_f f_n(x)\longrightarrow 0,\ 当\ n\to\infty. \tag{6·43}$$

记$K^*(u)=K^2(u)$. 易知，当K满足引理6·3的条件，则K^*亦然. 因而由引理6·3知

$$\lim_{n\to\infty}\frac{1}{h_n}\int K^*\left(\frac{u}{h_n}\right)f(x-u)du=f(x)\int K^*(u)du, \tag{6·44}$$

再由

$$\begin{aligned}\mathrm{Var}_f(f_n(x))&\leqslant\frac{1}{nh_n^2}E_f K^*\left(\frac{x-X_1}{h_n}\right)\\&=\frac{1}{nh_n}\cdot\frac{1}{h_n}\int K^*\left(\frac{u}{h_n}\right)f(x-u)du\end{aligned}$$

及$\lim\limits_{n\to\infty} nh_n=\infty$，即得(6·43). 定理证毕.

条件(6·41)的含义是：当$n\to\infty$时$h_n\to 0$，但其速度不能太快. 这与前一节的直观分析得到的结论是一致的. 下面讨论一致相合性.

定理6·3 设f一致连续，K为概率密度，且

(1) $K(u)$有可积的特征函数$k(u)$，

(2) $\lim\limits_{n\to\infty} h_n=0$, $\lim\limits_{n\to\infty} nh_n^2=\infty$.

则

$$\sup_x|f_n(x)-f(x)|\xrightarrow{p}0,\ 当\ n\to\infty. \tag{6·45}$$

证 由条件(1)及反演公式

$$K(u)=\frac{1}{2\pi}\int e^{-itu}k(t)dt$$

知$K(u)$有界，因而满足引理6·3的条件. 由定理6·1可得

$$\lim_{n\to\infty}\sup_x|Ef_n(x)-f(x)|=0.$$

记 $\varphi_n(u)=\frac{1}{n}\sum_{j=1}^n e^{iuX_j}=\int e^{iut}dF_n(t)$，其中 F_n 是 X_1，…，X_n 的经验分布函数。则有

$$f_n(x)=\frac{1}{h_n}\int K\left(\frac{x-t}{h_n}\right)dF_n(t)$$

$$=\frac{1}{2\pi h_n}\iint e^{-i\left(\frac{x-t}{h_n}\right)v}k(v)dvdF_n(t)$$

$$=\frac{1}{2\pi}\iint e^{-i(x-t)v}k(h_nv)dvdF_n(t)$$

$$=\frac{1}{2\pi}\int e^{-ixv}\phi_n(v)k(h_nv)dv.$$

于是

$$\sup_x|f_n(x)-Ef_n(x)|$$

$$\leqslant\frac{1}{2\pi}\int|k(h_nv)|\,|\varphi_n(v)-E\varphi_n(v)|dv$$

$$E\{\sup_x|f_n(x)-Ef_n(x)|\}$$

$$\leqslant\frac{1}{2\pi}\int|k(h_nv)|\sqrt{E|\varphi_n(v)-E\varphi_n(v)|^2}\,dv$$

$$\leqslant\frac{1}{2\pi}n^{-1/2}\int|k(h_nv)|dv$$

$$=\frac{1}{2\pi}\left(\frac{1}{nh_n^2}\right)^{1/2}\int|k(v)|dv\to0,\quad 当n\to\infty.$$

再由

$$E\{\sup_x|f_n(x)-f(x)|\}\leqslant E\{\sup_x|f_n(x)-Ef_n(x)|\}+\sup_x|Ef_n(x)-f(x)|,$$

即得

$$E\{\sup_x|f_n(x)-f(x)|\}\longrightarrow0,\quad 当\ n\to\infty.$$

由此即推出（6·45），定理证毕。

为了讨论强相合性，需要一个关于经验分布的概率不等式，由于该不等式的证明较为复杂，此处只叙述其结果而略去证明。

引理6·4 设 $X_1, \cdots, X_n$ 是来自连续分布函数 $F(x)$ 的独立同分布样本。$F_n(x)$ 是其经验分布函数。则存在绝对常数 $c>0$ 及 $0<\alpha\leqslant 2$ 使得：对任给 $\varepsilon>0$

$$P(\sup_x|F_n(x)-F(x)|\geqslant\varepsilon n^{-1/2})\leqslant c\exp(-\alpha\varepsilon^2). \qquad (6\cdot46)$$

定理6·4 设 K 是有界变差的概率密度，f 一致连续，若

$$\lim_{n\to\infty} h_n=0, \quad \lim_{n\to\infty} nh_n^2/(\log_n)=\infty, \qquad (6\cdot47)$$

则

$$\lim_{n\to\infty}\sup_x|f_n(x)-f(x)|=0, \text{ a.s.} \qquad (6\cdot48)$$

证 由定理 6·1 只须证：当 $n\to\infty$ 时

$$V_n\triangleq\sup_x|f_n(x)-Ef_n(x)|\longrightarrow 0, \text{ a.s.} \qquad (6.49)$$

由 K 有界变差，使用分部积分可知

$$\begin{aligned}V_n=&\sup_x\left|\frac{1}{h_n}\int K\left(\frac{x-y}{h_n}\right)dF_n(y)\right.\\&\left.-\frac{1}{h_n}\int K\left(\frac{x-y}{h_n}\right)dF(y)\right|\\=&\sup_x\left|\frac{1}{h_n}\int\left[F_n(y)-F(y)\right]dK\left(\frac{x-y}{h_n}\right)\right|\\\leqslant&V(k)\sup_y|F_n(y)-F(y)|/h_n,\end{aligned}$$

其中 $F(y)=\int_{-\infty}^{y}f(t)dt$，$V(K)$ 为 K 的全变差。因而对任给的 $\varepsilon>0$，

$$\begin{aligned}&P(\sup_x|f_n(x)-Ef_n(x)|\geqslant\varepsilon)\\&\leqslant P(\sup_x|F_n(x)-F(x)|\geqslant\varepsilon h_nV(k)^{-1})\\&\leqslant c\cdot\exp\{-\alpha n\varepsilon^2h_n^2V(K)^{-2}\}.\end{aligned}$$

由条件（6·47）即知

$$\sum_{n=1}^{\infty}c\exp\{-\alpha n\varepsilon^2h_n^2V(K)^{-2}\}<\infty,$$

依 Borel-Cantelli 引理即知 $V_n \to 0$, a.s. 当 $n\to\infty$. 定理证毕。

三、N.N. 估计的大样本性质

本段用 $\hat{f}_n(x)$ 表示由（6·6）所定义的 N.N. 估计。

定理6·5 设 $k=k_n$ 满足

$$k_n\to\infty,\quad k_n/n\to 0,\quad 当\ n\to\infty, \tag{6·50}$$

则当 $n\to\infty$时，

$$\hat{f}_n(x)\xrightarrow{P} f(x),\quad x\in c(f). \tag{6·51}$$

证 固定 $x\in c(f)$，对任给 $\varepsilon>0$ 有

$$P(|\hat{f}_n(x)-f(x)|>\varepsilon)=P\left(a_n(x)<\frac{k}{2n(f(x)+\varepsilon)}\right)+P\left(a_n(x)>\frac{k}{2n(f(x)-\varepsilon)}\right), \tag{6·52}$$

（若 $f(x)\leqslant\varepsilon$，第二项不出现）。记

$$p_n=\int_{x-\frac{k}{2n(f(x)+\varepsilon)}}^{x+\frac{k}{2n(f(x)-\varepsilon)}} f(t)\,dt,$$

y_n 为二项分布 $B(n,\ p_n)$ 变量。因 $x\in c(f)$，$\frac{k}{n}\to 0$，易见存在 $1>c>0$ 使得当 n 充分大时有

$$p_n\leqslant 2\frac{k}{2n(f(x)+\varepsilon)}c(f(x)+\varepsilon)=c\frac{k}{n}.$$

依 $a_n(x)$ 的定义可得

$$P\left(a_n(x)<\frac{k}{2n(f(x)+\varepsilon)}\right)=P(Y_n\geqslant k)$$

$$\leqslant P\left(\left|\frac{y_n}{n}-p_n\right|\geqslant(1-c)\frac{k}{n}\right)$$

$$\leqslant np_n(1-p_n)/(1-c)^2k^2$$

$$\leqslant c/[(1-c)^2k]\to 0,\quad 当\ n\to\infty.$$

用同样的方法可证：

$$P\left(a_n(x)>\frac{k}{2n(f(x)-\varepsilon)}\right)\longrightarrow 0,\quad 当\ n\to\infty.$$

由（6·52）即得证（6·51）．定理证毕。

上述定理的证明使用契比雪夫不等式，如改用 Hoeffding（1963）的一个较强的不等式，则可得到如下结果：

定理6·6 设 k_n 满足

$$k_n\to\infty,\ k_n/n\to 0,\ k_n/\log_n\to\infty \tag{6·53}$$

则有

$$\hat{f}_n(x)\to f(x),\ \text{a.s.}\ x\in c(f)，\text{当}\ n\to\infty。 \tag{6·54}$$

下面给出一个一致强相合的结果，

定理6·7 设 $k\triangleq k_n$ 满足

$$k_n\to\infty,\ k_n/n\to 0,\ k_n/\sqrt{n\log_n}\to\infty. \tag{6·55}$$

若 f 一致连续，则有

$$\lim_{n\to\infty}\sup_x|\hat{f}_n(x)-f(x)|=0,\ \text{a.s.} \tag{6·56}$$

证 对任给 $\varepsilon>0$

$$P(\sup_x|\hat{f}_n(x)-f(x)|>\varepsilon)$$

$$=P(\bigcup_x\{a_n(x)<k/2n(f(x)+\varepsilon)\})$$

$$+P(\bigcup_x\{a_n(x)>k/2n(f(x)-\varepsilon)\})\triangleq J_{n1}+J_{n2}.$$

由 f 的一致连续性，存在 $\delta>0$ 使得当 $|y-x|<\delta$ 时，就有 $|f(y)-f(x)|<\varepsilon/2$．记 $I_x=\left[x-\dfrac{k}{2n(f(x)+\varepsilon)},\ x+\dfrac{k}{2n(f(x)+\varepsilon)}\right]$ 是长度为 $d(x)=\dfrac{k}{n(f(x)+\varepsilon)}$ 的区间。显然当 n 充分大时，$d(x)\leqslant\dfrac{k}{n\varepsilon}<\delta$。记 F_n、F 所诱导的测度分别为 μ_n 及 μ。则当 n 充分大时

$$\mu(I_x)=\int_{I_x}f(t)\,dt\leqslant\frac{k}{n(f(x)+\varepsilon)}(f(x)+\varepsilon/2).$$

因而当 n 充分大时

$$\left\{a_n(x)<\frac{k}{2n(f(x)+\varepsilon)}\right\}\subset\left\{\mu_n(I_x)\geqslant\frac{k}{n}\right\}$$

$$\subset \left\{ \mu_n(I_x) - \mu(I_x) \geqslant \frac{k\varepsilon}{2n(f(x)+\varepsilon)} \right\}$$

$$\subset \left\{ |\mu_n(I_x) - \mu(I_x)| \geqslant \frac{k\varepsilon}{2n(M+\varepsilon)} \right\},$$

其中 $M = \sup\limits_x f(x)$．记 $c_0 = \varepsilon / 4(M+\varepsilon)$，则当 n 充分大时

$$J_{n1} \leqslant P\left(\bigcup_x \left\{ |\mu_n(I_x) - \mu(I_x)| \geqslant 2c_0 \frac{k}{n} \right\}\right)$$

$$\leqslant P\left(\sup_x |F_n(x) - F(x)| \geqslant c_0 \frac{k}{n}\right).$$

由引理 6·4，存在绝对常数 $c>0$，$c_1>0$ 使得当 n 充分大时

$$J_{n1} \leqslant c\exp(-c_1 k^2/n).$$

由条件（6·55）即知 $\sum\limits_{n=1}^{\infty} J_{n1} < \infty$．同理可证 $\sum\limits_{n=1}^{\infty} J_{n2} < \infty$．因而由 Borel-Cantelli 引理即知（6·56）成立，定理证毕。

至于 N.N. 估计的均方相合性，其证明比较复杂。有兴趣的读者可参看有关文献。另一个有关问题，即从大样本角度比较 N.N. 估计与 Rosenblatt 估计的优劣。由于这一问题过于专门，在此略去。

四、高维情形

到此为止，我们讨论的密度估计都是基于一维数据。这是由于从理论分析角度，一维情形简单明瞭，且不少的大样本结果在高维都有类似的推广。然而，密度估计的不少重要应用领域，涉及的是高维数据。无论从应用还是理论分析，高维情形都有其特殊性。本段仅就高维情形的某些方面作一点简单的注解。

下面均设 $X_1, \cdots, X_n$ 是来自未知 d 维密度 $f(x)$ 的独立同分布样本。

1．光滑参数的设计

我们可将一元核估计的定义推广为：

$$f_n(x) = \frac{1}{nh_n^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right) \tag{6·57}$$

其中 $K(\cdot)$ 是 R^d 中的密度函数，例如 d 维标准正态密度，$h_n>0$ 是窗宽。这是通常使用的高维核估计的定义。在这一定义中，对数据的每一分量用同一刻度因子 h_n 加以光滑。当数据点在某一方向上的变异比其它方向要显著地大时，这一定义明显地不合适。在这种情况下，不如使用一个常向量或常数矩阵作为光滑参数来得好。另一种方法是先将数据作刻度变换，以降低数据点的各向变异，再对经处理的数据使用定义（6·57）。Fukunaga（1972）曾提出如下的变换方法：记 S 为 $X_1, \cdots, X_n$ 的样本协差阵，作变换 $\widetilde{X}_i=S^{-1/2}X_i$，$i=1, 2, \cdots, n$。然后使用一个径向对称核 K 加以光滑，最后变回原数据。如此得到的估计可以表成

$$f_n(x)=\frac{(\det S)^{-1/2}}{nh_n^d}\sum_{i=1}^{n}k[h_n^{-2}(x-X_i)'S^{-1}(x-X_i)], \tag{6·58}$$

其中

$$k(x'x)=K(x).$$

这样作的好处是：变换后的数据 $\widetilde{X}_1, \cdots, \widetilde{X}_n$ 其样本协差阵是单位阵，因而消除了数据的各向变异的差别。但公式（6·58）的计算量较大。文献中其它的讨论还很多，这里不一一列举。

2．尾部估计

一般说来，在低维情形 f 的尾部估计失当影响不大。这是因落在尾部区域中的数据很少。故而绝大部分样本可看成来自截尾分布。然而当维数 d 增大时，情况就有明显的差别。例如 $d=10$，f 为标准正态密度。则平均来说，大约有过半的数据落在区域

$$D\triangleq\{x: f(x)<(0.01)\times f(o)\}$$

之中。事实上，如记 $f(X_1)/f(o)$ 的中位数为 m，则由 $f(X_1)/f(o)=\exp\left\{-\frac{1}{2}X_1'X_1\right\}\sim\exp\left(-\frac{1}{2}\chi_{10}^2\right)$，其中 χ_{10}^2 为自由度为 10 的 χ^2 分布。而 χ_{10}^2 的中位数为 9.34，可得 $m=\exp(-9.34/2)=0.0094<0.01$．因而

$$\frac{1}{2}=P(f(X_1)/f(0)\leqslant m)\leqslant P(f(X_1)/f(0)<0.01).$$

于是

$$E[{}^{*}\{i:X_i\in D,\ i=1,\ 2,\ \cdots,\ n\}]$$
$$=nP(f(X_1)/f(0)<0.01)\geqslant\frac{n}{2}.$$

此例也表明：与低维情形相反，低密度区域是高维分布的非常重要部分。因而在高维情形，对 f 的尾部估计需要十分小心。

3．对给定估计精度，维数对最低限度的样本容量的影响。

我们以均方误差作为精度的测度，则对给定的精度及假设理论分布，原则上可定出一个最低限度的样本容量。在实际问题中，当然希望这个值越小越好、但是随着维数的增大，最低样本容量的增大是非常之快。例如 f_n 是正态核密度估计，f 是标准正态密度。考察 $x=0$ 处的均方误差。如要求

$$E(f_n(0)-f(0))^2\leqslant(0.1)\times f(0)^2,$$

则当 $d=2$ 时，$n=19$；$d=3$ 时，$n=67$；而当 $d=10$ 时，$n=842000$. 如此之大的样本容量，在实际问题中是无法承受的。因而根据实际需要，不断改进估计方法是个重要课题。

§6·3 非参数回归

一、引言

设在一实际问题中，我们感兴趣的变量X与Y（均可为多维）有某种相关关系。即当给定$X=x$时，虽然还不足以确定Y的值，但Y的条件分布由x所确定，为方便计，称X为自变量，Y为因变量。例如X是某种农作物单位面积的施肥量和播种量，此时X为二维的自变量，而Y为该作物的亩产量，Y的值当然同X之取值有关，但还未达到由它所完全确定的程度，因为Y还受到诸如管理水平、气候变化及其他大量因素的影响。但在许多实际

问题中，可以认为当X取一定值时，能确定Y的条件分布，后者对X取值的依赖关系，即是最广意义下的回归关系。在经典回归分析中，常假定（X'，Y'）$'$有多元正态分布$N(\mu, \Sigma)$，其中

$$\mu=\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix},\quad \Sigma=\begin{pmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{pmatrix} \tag{6·59}$$

μ，Σ表达式中的分块相应于X，Y的维数。在此假定下，当给定$X=x$时，Y的条件分布仍为多元正态．Y的条件期望为

$$m(x)\triangleq E(Y|X=x)=E(Y|x)=\mu_2+A_{21}A_{11}^{-1}(x-\mu_1). \tag{6·60}$$

函数$m(x)$常称为（Y对X的）回归函数，它描述了Y的条件期望随X值变化的情况。若有来自（X，Y）的随机样本$Z_n=\{(X_1, Y_1),\cdots(X_n, Y_n)\}$，则可基于$Z_n$，作出（6·60）中未知参数的最小二乘估计．理论和实践都证明了在上述正态回归模型下，最小二乘估计有种种优良性质．然而在很多实际问题中，正态性不一定成立，这时我们要另找办法去估计回归函数$E(Y|x)$及其他有意义的量，如条件方差$\mathrm{Var}(Y|x)$等。有时可通过直接估计Y（在给定$X=x$之下）的条件分布来达到这一目的．例如，考虑下面这样一种情况：对给定的x，在X_1，…，X_n中有若干个（数目较大）X_i恰好等于x．譬如，设$X_{i_j}=x$，$j=1$，2，…，k，则可用

$$F_n(y|x)\triangleq\frac{1}{k}\sum_{j=1}^{k} I[Y_{i_j}\leqslant y]$$

来估计给定$X=x$时，Y的条件分布函数$F(y|x)$，然后，用

$$\hat{m}_n(x)\triangleq\int y dF_n(y|x)=\frac{1}{k}\sum_{j=1}^{k} Y_{i_j} \tag{6·61}$$

作为回归函数$m(x)$的估计．这种作法只是在很特殊的情况下才可行，一般说来，对给定的x，可能在X_1，…，X_n中有很少的样本（甚至一个也没有）恰好等于x，上述作法就行不通了，因而，必须寻找一种普遍适用的估计条件分布的方法．我们还可从另一角度考察这一问题．如能找到一种方法估计$E(f(y)|x)$，

其中 f 为任一实函数，则当 $f(Y)=I[Y\in A]$（A 是某个区间）时，就能估计条件概率；而当 $f(Y)=Y$ 或 $f(Y)=[Y-E(Y|x)]^2$ 时，即得条件均值及条件方差的估计。于是诸多条件量的估计问题可以归结成估计回归函数 $E(Z|x)$ 的问题，其中 $Z=f(Y)$。仍用 Y 代替记号 Z，我们可以将问题一般地表述为：设有因变量 Y（为一维）与自变量 X（d 维）配对，(X, Y) 的分布未知，只假定 $E|Y|<\infty$。今有来自 (X, Y) 的随机样本 (X_i, Y_i)，$i=1, 2, \cdots, n$，要求基于该样本估计回归函数 $m(x)=E(Y|x)$，即构造估计 $m_n(x)=m_n(x; X_1, Y_1, \cdots, X_n, Y_n)$，使得对每一个 $x\in R^d$，用 $m_n(x)$ 作 $m(x)$ 的估计。

Stone 在 1977 年提出了一种非参数回归估计的权函数方法，并在理论上论证了这种方法的优良性（主要是其大样本性质）。Stone 的方法引起了广泛的重视。在这段时间内，这一方向取得了很大进展。本节着重介绍权函数方法的有关概念以及方法的应用，而对有关理论结果，只作必要的简单介绍.

二、权函数法

我们从上面提到的特殊情形出发。设有 Y 与 X 配对，(X_i, Y_i) $i=1, 2, \cdots, n$ 是来自 (X, Y) 的随机样本。对给定的 $x\in R^d$，将 $X_1, \cdots, X_n$ 中恰好等于 x 的那些样本挑选出来。例如其下标为 $i_1, i_2, \cdots, i_k$（显然 $i_1, i_2, \cdots, i_k$ 既同 x 有关，也同样本 $X_1, \cdots, X_n$ 有关），则在估计 $m(x)$ 时，样本 (X_{i_j}, Y_{i_j})（$j=1,2, \cdots, k$）显得比别的样本重要。如用 $W_{ni}(x)\triangleq W_{ni}(x; X_1, \cdots, X_n)$ 表示样本 (X_i, Y_i) 在估计 $m(x)$ 时的重要程度，或者说样本 (X_i, Y_i) 的权，则 $W_{ni}(x)$ 应有如下形式：

$$W_{ni}(x)=\begin{cases}\dfrac{1}{k}, & \text{当 } i \text{ 是 } i_1, i_2, \cdots, i_k \text{ 之一}\\ 0, & \text{对别的 } i\end{cases} \qquad (6\cdot62)$$

这是因 $X_{i_j}=x$, $j=1,2,\cdots k$。因而(X_{i_j}, Y_{i_j})应有相同的权，而总数一共为k个，因此（6·62）的结构是合理的。由此（6·61）可改写为

$$\hat{m}_n(x)=\sum_{j=1}^{n} W_{nj}(x) Y_j。$$

将上述构造过程加以推广，就得出如下的一般定义。

定义6·6 以n记样本大小，则n个形如$W_{ni}(x)=W_{ni}(x; X_1, \cdots, X_n)$ $(i=1, 2, \cdots, n)$的函数，称为权函数（权函数可以指这n个的整体，或者其中任一个），又若

$$W_{ni}(x)\geqslant 0,\ 1\leqslant i\leqslant n;\ \sum_{i=1}^{n} W_{ni}(x)=1, \qquad (6\cdot 63)$$

则称$\{W_{ni}\}$为概率权函数。对给定的权函数$\{W_{ni}\}$，定义回归函数$m(x)$的估计为

$$m_n(x)=\sum_{i=1}^{n} W_{ni}(x) Y_i, \qquad (6\cdot 64)$$

并称$m_n(x)$为$m(x)$的一个权函数估计。

从（6·64）可以看出：一个权函数估计完全由给定的权函数$\{W_{ni}\}$所确定，而权函数的分布只同X的分布有关。样本Y_i（或者说(X_i, Y_i)）对$m_n(x)$的贡献，除其本身之值外，还取决于权W_{ni}。因而权$W_{ni}(x)$表示在估计$m(x)$时，样本(X_i, Y_i)所起的作用的"大小"。由上述定义可知，（6·62）式所确定的$\{W_{ni}\}$是一种特殊的概率权函数，而估计（6·61）是由之确定的权函数估计。下面将介绍两种构造权函数的方法，即近邻权及核权方法。

1。近邻权方法

其直观想法是，对给定的样本$X_1, \cdots, X_n$及$x\in \mathrm{R}^d$，虽然可能没有一个X_i恰好等于x，但可将"等于x"的要求降低为"与x接近"。依每个X_i对给定x的距离重新排序，与x距离越近的其重要程度越大。为了简便起见，我们选用欧氏距离$\|\cdot\|$，将样本$X_1, \cdots, X_n$依在距离$\|\cdot\|$的意义下，与x的接近程度排

$$\|X_{R_1}-x\|\leqslant\|X_{R_2}-x\|\leqslant\cdots\leqslant\|X_{R_n}-x\|. \qquad (6\cdot65)$$

再选定 n 个常数 C_{n1}, C_{n2}, …, C_{nn}, 满足条件

$$C_{n1}\geqslant C_{n2}\geqslant\cdots\geqslant C_{nn}\geqslant0,\quad \sum_{i=1}^{n}C_{ni}=1, \qquad (6\cdot66)$$

用$\{C_{ni}\}$作为权的大小的计量。因 X_{R_1} 与 x 最接近,赋予权 C_{n1},其次一个是 X_{R_2}, 赋予权 C_{n2}, …等等。最后定义权函数为

$$W_{nR_i}(x)=C_{ni},\quad i=1,\ 2,\ \cdots,\ n \qquad (6\cdot67)$$

当(6·65)中有等号出现时, 可采用"足标靠前原则", 即若有 $1\leqslant i<j\leqslant n$, 使$\|X_i-x\|=\|X_j-x\|$, 则在(6·65)的排序中, X_i 出现在 X_j 之前。称如此定义的权函数为近邻权函数。注意到(6·65)中的下标 R_1, R_2, …, R_n 既同 x 有关, 又同样本 X_1, …, X_n 有关。不难验证近邻权是概率权函数。由此定义可知(6·62)所确定的权函数是近邻权的一个特例,在那里有 $C_{n1}=\cdots=C_{nk}=\frac{1}{k}$, $C_{ni}=0$, 当 $k+1\leqslant i\leqslant n$; 而 $R_j=i_j$, $j=1,\ 2,\ \cdots,\ k$ (若$1\leqslant i_1<i_2<\cdots<i_k\leqslant n$)。近邻权方法在理论上已经证明了有诸多优良的大样本性质, 但是其计算较复杂, 对每一个 x, 要重新按(6·65)排序。另外,在近邻权的定义中距离$\|\cdot\|$与(6·66)中的常数$\{C_{ni}\}$都有很大的选择余地, 这正如在核密度估计中, 有一个核与窗宽 h_n 的选择问题。

2. 核函数法

选定 $\mathbb{R}^d$ 上的核函数 $K(\cdot)$ 及窗宽 h_n, 然后定义

$$W_{ni}(x)=K\left(\frac{x-X_i}{h_n}\right)\Big/\sum_{j=1}^{n}K\left(\frac{x-X_j}{h_n}\right),\quad i=1,\ \cdots,\ n, \qquad (6\cdot68)$$

称此$\{W_{ni}\}$为核权函数。由权函数的定义可知, 核权函数也是概率权函数, 相应的权函数估计为

$$m_n(x)=\sum_{i=1}^{n}K\left(\frac{x-X_i}{h_n}\right)Y_i\Big/\sum_{j=1}^{n}K\left(\frac{x-X_j}{h_n}\right).$$

$$(6 \cdot 69)$$

估计（6·69）的合理性可作如下解释：设（X，Y）有联合密度 $f(x, y)$ 则有

$$m(x) = E(y|x) = \int yf(x, y)dy \Big/ \int f(x, y)dy$$
$$\triangleq \int yf(x, y)dy / f_X(x).$$

边缘密度 $f_X(x)$ 的核估计为 $\frac{1}{nh_n^d}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)$，而 $\int yf(x, y)dy$ 可用 $\frac{1}{nh_n^d}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)Y_i$ 去估计。分别以这两个估计作分母和分子即得（6·69）。有趣的是，当取

$$K(u) = \begin{cases} \frac{1}{C_d}, & \text{当} \|u\| \leqslant 1 \\ 0, & \text{对其它 } u \end{cases}$$

其中 C_d 为 R^d 中单位球的体积，记

$$B(x, a) = \{t: t \in \mathrm{R}^d, \ \|t - x\| < a\},$$

则有

$$W_{ni}(x) = \begin{cases} 1/\{X_1, \cdots, X_n \text{落在球} B(x, h_n) \text{的个数}\}, & \text{当} X_i \in B(x, h_n) \\ 0, & \text{当} X_i \overline{\in} B(x, h_n) \end{cases}$$

又回到近邻权的情况。核权函数的优点是有一个明确的关于 x 的统一的表达式，从而便于计算。但由于（6·69）的分母是随机变量，给理论处理带来一定的困难。

三、权函数估计的相合性

同概率密度估计一样，非参数回归估计的理论分析，到目前为止其深入的结果也只在大样本方面。本段着重介绍由 Stone 首先提出的权函数估计的矩相合性。可以说在这方面的几乎所有结果论证都较复杂。因此我们只着重有关概念的阐述，而对提到的少数几个定理，其证明都省略。本段将采用以下记号

$W_{ni}(X) \triangleq W_{ni}(x)|_{x=X}$，$m_n(X) \triangleq m_n(x)|_{x=X}$等等。

设已给定权函数$\{W_{ni}\}$，将任意Y与X配对，考虑$m(x)=E(Y|x)$的估计。依（6·64）应从$m_n(x)=\sum_{i=1}^{n} W_{ni}(x)Y_i$估计之，其中$Y_1$，$Y_2$，…，$Y_n$是来自给定$Y$的样本，其绝对偏差

$$d_n(x) \triangleq |m_n(x)-m(x)|$$

可以用来衡量权函数$\{W_{ni}\}$（在逐点意义下）的优劣。文献上有讨论权函数估计的逐点相合性及逐点强相合性，其含义是指分别具有下述性质：

$$d_n(x) \xrightarrow{P} 0, \text{ 当 } n\to\infty$$

及

$$d_n(x) \longrightarrow 0, \text{ a.s.}, \text{ 当 } n\to\infty.$$

另外一种途径则是考虑整体精度，即绝对偏差$d_n(X)$的平均。直观上看，一个好的权函数$\{W_{ni}\}$应有如下性质：不论与X配对的Y如何选取（当然要满足某些最必要的条件，例如对Y有一定阶的矩），当$n\to\infty$时，由之产生的$d_n(X)$，"平均地说"应收敛于0。这一想法导致"矩相合"的概念。这是Stone首先提出的，其确切定义如下：

定义6·7 设$\{W_{ni}\}$是给定的权函数，若对任意的$r\geqslant 1$及任一满足

$$E|Y|^r<\infty \qquad (6\cdot 70)$$

的Y，都有

$$\lim_{n\to\infty} E(d_n(X))^r=0, \qquad (6\cdot 71)$$

则称$\{W_{ni}\}$为矩相合的。

注意，此定义指权函数本身的相合性而不直接指权函数估计的相合性。这是因为定义中的条件满足与否，只取决于权函数本身。还值得注意的是，在此定义中要求（6·71）对所有$r\geqslant 1$都成立。因而矩相合与通常的r阶矩（对某个r）相合不同。关于矩相合的一个基本结果是如下的

定理6·8 设 $\{W_{ni}\}$ 为给定的概率权函数，则其矩相合的充要条件是

(1) 存在有限常数 C，使得对任一非负函数 f 都有

$$E\left(\sum_{i=1}^{n} W_{ni}(X) f(X_i)\right) \leqslant CEf(X);$$

(2) 对任给 $\varepsilon>0$，当 $n\to\infty$ 时有

$$\sum_{i=1}^{n} W_{ni}(X) I[\|X_i - X\| > \varepsilon] \xrightarrow{P} 0;$$

(3) $\max\limits_{1\leqslant i\leqslant n} W_{ni}(X) \xrightarrow{P} 0$.

这一定理的条件（1）较难验证，它是一个纯技术性条件，不易作出直观上的解释。今只对条件(2)、(3)作一直观说明。条件(2)可理解为对于与 x 距离超过某种限度的那些样本 X_i，其权的总和很小，因而在估计 $m(x)$ 时，主要依据最接近 x（即在此限度以内）的那些样本。条件(3)意味着，作为单独的一个样本点 X_i，不论它与 x 的距离多么接近，所起的作用总是很小的。这正如概率论中的中心极限定理，单个样本的作用小，但其总和随着 n 增大，其作用也随之增大。这些要求是与构造权函数的基本思想一致，因而是合理的。下面是这一基本定理对近邻权及核权函数的应用。

定理6·9 设给定常数序列 $\{C_{ni}\}$ 满足（6·66），而 $\{W_{ni}\}$ 是由（3·65）及（6·67）所确定的近邻权。如

(1) $\lim\limits_{n\to\infty} \sum\limits_{\varepsilon n<i\leqslant n} C_{ni}=0$，对任何 $\varepsilon>0$；

(2) $\lim\limits_{n\to\infty} C_{n1}=0$，

则 $\{W_{ni}\}$ 为矩相合。

此处的条件(2)显然能推出定理 6·8 的条件(3)，又以概率1对每一 $x\in R^d$，$\varepsilon>0$，存在 $\eta>0$ 使得

$$\begin{aligned}\sum_{i=1}^{n} W_{ni}(x)\, I_{[\|X_i-x\|>\varepsilon]} &= \sum_{i=1}^{n} W_{nR_i}(x)\, I_{[\|X_{R_i}-x\|>\varepsilon]} \\ &= \sum_{i\geqslant \eta n} C_{ni},\end{aligned}$$

因而此处的条件(1)可推出定理 6·8 之条件(2)。但定理 6·8 之条件（1）的验证较为复杂。

定理6·10 设$\{W_{ni}\}$为以K为核的核权函数，而K为$\mathbb{R}^d$上具有紧支撑的有界概率密度。若

$$h_n \to 0, \quad nh_n^d \to \infty, \quad 当\ n \to \infty,$$

则$\{W_{ni}\}$为矩相合的。

这一结果是由 Devroye 和 Wagner 得到的，其缺点是对核的要求过严。

四、应用

本段介绍权函数估计的若干应用。

1。条件二阶矩估计

设有 q 维变量 $Y=(Y^{(1)}, \cdots, Y^{(q)})$ 与X配对，而(X_i, Y_i)，$i=1, \cdots, n$ 是来自(X, Y)的随机样本，且已给定了权函数$\{W_{ni}\}$，要求估计给定 $X=x$ 时，Y的条件二阶矩。例如Y的分量的条件方差、条件协方差及条件相关系数。因

$$\mathrm{Var}(Y^{(i)}|x)=E[(Y^{(i)})^2|x]-[E(Y^{(i)}|x)]^2,$$

$$\mathrm{Cov}(Y^{(i)}, Y^{(j)}|x)=E[Y^{(i)}Y^{(j)}|x] -E[Y^{(i)}|x]E[Y^{(j)}|x],$$

$$\rho(Y^{(i)}, Y^{(j)}|x)=\mathrm{Cov}(Y^{(i)}, Y^{(j)}|x)/\sqrt{\mathrm{Var}(Y^{(i)}|x)\mathrm{Var}(Y^{(j)}|x)}$$

$$(i, j=1, 2, \cdots, q)$$

只须估计条件协方差。记估计量为$\mathrm{Cov}_n(Y^{(i)}, Y^{(j)}|x)$。由权函数估计的定义，分别构造对子$(X, Y^{(i)})$，$(X, Y^{(j)})$及$(X, Y^{(i)}Y^{(j)})$所产生的回归函数的估计，再依协方差结构可得

$$\mathrm{Cov}_n(Y^{(i)}, Y^{(j)}|x)=\sum_{k=1}^{n} W_{nk}(x)Y_k^{(i)}Y_k^{(j)} -\sum_{k=1}^{n} W_{nk}(x)Y_k^{(i)} \sum_{k=1}^{n} W_{nk}(x)Y_k^{(j)}, \tag{6·72}$$

再令 $j=i$，又得 $\mathrm{Var}(Y^{(i)}|x)$ 的估计 $\mathrm{Var}_n(Y^{(i)}|x)$。最后定义 $\rho(Y^{(i)}, Y^{(j)}|x)$ 的估计为

$$\rho_n(Y^{(i)}, Y^{(j)}|x)=\mathrm{Cov}_n(Y^{(i)}, Y^{(j)}|x)/\sqrt{\mathrm{Var}_n(Y^{(i)}|x)\,\mathrm{Var}(Y^{(j)}|x)}, \tag{6·73}$$

当 $\{W_{ni}\}$ 是概率权时，由

$$\left(\sum_{k=1}^{n} W_{nk}(x)Y_k^{(i)}\right)^2\leqslant\left(\sum_{k=1}^{n} W_{nk}(x)\right)\left[\sum_{k=1}^{n} W_{nk}(x)(Y_k^{(i)})^2\right]=\sum_{k=1}^{n} W_{nk}(x)(Y_k^{(i)})^2$$

知，估计 $\mathrm{Var}_n(Y^{(i)}|x)\geqslant 0$。同样可知 $|\rho_n(Y^{(i)}, Y^{(j)}|x)|\leqslant 1$。因而对于条件二阶矩估计，要求 $\{W_{ni}\}$ 是概率权函数是合理的。对于上述估计，我们有如下的大样本性质。

定理6·11 设 $\{W_{ni}\}$ 为相合的概率权函数，且 $E\|Y\|^2<\infty$，则有

$$\lim_{n\to\infty} E\{|\mathrm{Cov}_n(Y^{(i)}, Y^{(j)}|X)-\mathrm{Cov}(Y^{(i)}, Y^{(j)}|X)|\}=0. \tag{6·74}$$

又若以概率1有 $\mathrm{Var}(Y^{(i)}|X)>0$，$\mathrm{Var}(Y^{(j)}|X)>0$，则对任给 $r>0$，有

$$\lim_{n\to\infty} E\{|\rho_n(Y^{(i)}, Y^{(j)}|X)-\rho(Y^{(i)}, Y^{(j)}|X)|^r\}=0. \tag{6·75}$$

证 先证明两点预备事实。

(1) 设 ξ_n，η_n，ξ，η 都是随机变量，若

$$\lim_{n\to\infty} E(\xi_n-\xi)^2=\lim_{n\to\infty} E(\eta_n-\eta)^2=0,$$

则

$$\lim_{n\to\infty} E|\xi_n\eta_n-\xi\eta|=0.$$

事实上，由假设知存在常数 $M<\infty$ 使得 $E\xi_n^2\leqslant M$。注意到 $|\xi_n\eta_n-\xi\eta|\leqslant|\xi_n-\xi||\eta|+|\eta_n-\eta||\xi_n|$，由 Cauchy-Schwartz

不等式得到

$$(E|\xi_n-\xi||\eta|)^2\leqslant E|\xi_n-\xi|^2E|\eta|^2\to 0,\ \text{当}\ n\to\infty,$$

$$(E|\eta_n-\eta||\xi_n|)^2\leqslant ME|\eta_n-\eta|^2\to 0,\ \text{当}\ n\to\infty,$$

因而 $E|\xi_n\eta_n-\xi\eta|\to 0$，当 $n\to\infty$.

(2) 设 $\xi_n\geqslant 0$，$\xi\geqslant 0$，且 $\lim\limits_{n\to\infty}E|\xi_n-\xi|=0$，

则

$$\lim_{n\to\infty}E(\sqrt{\xi_n}-\sqrt{\xi})^2=0.$$

事实上，$|\sqrt{\xi_n}-\sqrt{\xi}|\leqslant\sqrt{\xi_n}+\sqrt{\xi}$，从而由

$$(\sqrt{\xi_n}-\sqrt{\xi})^2\leqslant|\sqrt{\xi_n}+\sqrt{\xi}||\sqrt{\xi_n}-\sqrt{\xi}|=|\xi_n-\xi|,$$

即得所证.

现回到定理的证明. 假设 $E|Y^{(i)}|^2<\infty$，$1\leqslant i\leqslant q$. 从而 $E|Y^{(i)}Y^{(j)}|<\infty$。由矩相合定义，分别考虑 $Y^{(i)}Y^{(j)}$ 与 X 及 $Y^{(i)}$ 与 X 配对，并分别取 $r=1$ 及 2，得到

$$\lim_{n\to\infty}E\Big|\sum_{k=1}^{n}W_{nk}(X)Y_k^{(i)}Y_k^{(j)}-E(Y^{(i)}Y^{(j)}|X)\Big|=0, \tag{6·76}$$

$$\lim_{n\to\infty}E\Big|\sum_{k=1}^{n}W_{nk}(X)Y_k^{(i)}-E(Y^{(i)}|X)\Big|^2=0. \tag{6·77}$$

在预备事实(1)中，取 $\xi_n=\sum\limits_{k=1}^{n}W_{nk}(X)Y_k^{(i)}$，$\eta_n=\sum\limits_{k=1}^{n}W_{nk}(X)Y_k^{(j)}$，$\xi=E(Y^{(i)}|X)$，$\eta=E(Y^{(j)}|X)$。则由(6·77)知，

$$E(\xi_n-\xi)^2\to 0,\ E(\eta_n-\eta)^2\to 0,\ \text{当}\ n\to\infty.$$

因而由预备事实(1)即得

$$\lim_{n\to\infty}E\Big|\Big(\sum_{k=1}^{n}W_{nk}(X)Y_k^{(i)}\Big)\Big(\sum_{k=1}^{n}W_{nk}(X)Y_k^{(j)}\Big)-E(Y^{(i)}|X)E(Y^{(j)}|X)\Big|=0. \tag{6·78}$$

由(6·76)、(6·78)即得

$$\lim_{n\to\infty}E|\mathrm{Cov}_n(Y^{(i)},Y^{(j)}|X)-\mathrm{Cov}(Y^{(i)},Y^{(j)}|X)|=0,$$

$$i,j=1,2,\cdots,q.$$

此即（6·74）成立。令 $j=i$，又得

$$\lim_{n\to\infty} E|\mathrm{Var}_n(Y^{(i)}|X)-\mathrm{Var}(Y^{(i)}|X)|=0.$$

再由预备事实(2)，有

$$\lim_{n\to\infty} E|\sqrt{\mathrm{Var}_n(Y^{(i)}|X)}-\sqrt{\mathrm{Var}(Y^{(i)}|X)}|=0,$$

$$i=1, \cdots, q. \qquad (6\cdot79)$$

再由已证的（6·74）及（6·73），并注意到

$$\mathrm{Var}(Y^{(i)}|X)>0, \text{ a. s.}, \quad \mathrm{Var}(Y^{(j)}|X>0, \text{ a. s.},$$

可得

$$\rho_n(Y^{(i)}, Y^{(j)}|X) \xrightarrow{p} \rho(Y^{(i)}, Y^{(j)}|X), \text{当 } n\to\infty。$$

但 $|\rho_n(Y^{(i)}, Y^{(j)}|X)|\leqslant1$，$|\rho(Y^{(i)}, Y^{(j)}|X)|\leqslant1$，a.s.由此易得（6·75），定理证毕。

2．条件分位数估计

设 Z 为任一随机变量，$p\in(0, 1)$。如果实数 ζ_p 满足

$$P(Z<\xi_p)\leqslant p\leqslant P(Z\leqslant\xi_p)$$

则称 ξ_p 为 Z 的 p 分位数。若 F 为 Z 的分布函数，有时也称 ξ_p 为 F 的 p 分位数。一般 ξ_p 并不唯一，但有如下性质：令

$$c=\sup\{t:F(t)\leqslant p\}, \quad d=\inf\{t:F(t)\geqslant p\}, \qquad (6\cdot80)$$

则 $-\infty<d\leqslant c<\infty$，且 ξ_p 为F的 p 分位数当且仅当 $\xi_p\in[c, d]$。今设有随机变量Y与X配对（X仍设为 d 维向量），以$F(\cdot|x)$ 记给定 $X=x$ 时，Y的条件分布函数，记其 p 分位数为 $\xi(p|x)$。(X_i, Y_i)，$i=1, \cdots, n$ 是来自(X, Y)的随机样本，$\{W_{ni}\}$为基于 $X_1, \cdots, X_n$ 的一个给定的权函数。则 $F(\cdot|x)$ 的权函数估计为

$$F_n(y|x)=\sum_{i=1}^{n} W_{ni}(x) I_{(Y_i<y)}. \qquad (6\cdot81)$$

易知若$\{W_{ni}\}$为概率权函数，则 $F_n(y|x)$ 作为 y 的函数是一维分布函数。以 $F_n(\cdot|x)$ 的任一p 分位数 $\xi_n(p|x)$ 作为 $\xi(p|x)$ 的估计。显然 $\xi_n(p|x)$ 由权函数$\{W_{ni}\}$所确定，但并不唯一。然而出乎意料的是：在某种限制下，当 $\xi(p|x)$ 唯一时(此

时 $\xi_n(p|x)$ 仍不必唯一），不论如何选择 $\xi_n(p|x)$，其大样本极限是唯一的。实际上，若记 $L(p|x)$，$U(p|x)$ 分别为 $F(\cdot|x)$ 的 p 分位区间的左、右端点，而 $L_n(p|x)$，$U_n(p|x)$ 为 $F_n(\cdot|x)$ 的 p 分位区间的左、右端点，则 $[L_n(p|x), U_n(p|x)]$ 可以作为 $\xi(p|x)$ 的一个区间估计。上面提到的那个事实可以看作是 $\xi(p|x)$ 的区间估计的一项渐近性质。我们将此事实确切表示为下述的定理。

定理6·12 设 $\{W_{ni}\}$ 为矩相合的概率权函数，若 $F(\cdot|X)$ 以概率 1 有唯一的 p 分位数 $\xi(p|X)$，则不论如何选择 $\xi_n(p|x)$，都有

$$\xi_n(p|X) \xrightarrow{P} \xi(p|X). \qquad (6\cdot82)$$

又若对某个 $r>0$ 有 $E|Y|^r<\infty$，则

$$\lim_{n\to\infty} E[|\xi_n(p|x)-\xi(p|x)|^r]=0. \qquad (6\cdot83)$$

证明：依假设，以概率为 1 地有

$$L(p|X)=U(p|X)=\xi(p|X),$$

因而有

$$|\xi_n(p|X)-\xi(p|X)|\leqslant[U_n(p|X)-U(p|X)]^+ + [L_n(p|X)-L(p|X)]^-, \text{ a.s.}$$

为此只须证：当 $n\to\infty$ 时有

$$[U_n(p|X)-U(p|X)]^+ \xrightarrow{P} 0, \qquad (6\cdot84)$$

$$[L_n(p|X)-L(p|X)]^- \xrightarrow{P} 0, \qquad (6\cdot85)$$

以及

$$E\{[U_n(p|X)-U(p|X)^+\}^r \longrightarrow 0, \qquad (6\cdot86)$$

$$E\{[L_n(p|X)-L(p|X)]^-\}^r \longrightarrow 0, \qquad (6\cdot87)$$

而 (6·84)、(6·85) 包含在下述结论之中：

对每 $\varepsilon>0$，$\lim\limits_{n\to\infty} P\{L_n(p|X)\geqslant L(p|X)-\varepsilon\}=1$，

$$\lim_{n\to\infty} P\{U_n(p|X)\leqslant U(p|X)+\varepsilon\}=1.$$

其思想是充分利用矩相合的概念，构造合适的因变量与X配对。

为了证（6·86）及（6·87），必须证 $\{|L_n(p|X)|^r\}$ 以及 $\{|U_n(p|X)|^r\}$一致可积，以及 $L(p|X)$，$U(p|X)$ r 阶矩有限，后者可由 $E|Y|^r<\infty$ 推出，而前者通过建立下述估计得到

$$E\{|L_n(p|X)|^r I_{[|L_n(p|X)>M]}\}\leqslant C_p E(|Y|^r I_{[|Y|>M]}),$$

$$E\{|U_n(p|X)|^r I_{[|U_n(p|X)<M]}\}\leqslant C_p E(|Y|^r I_{[|Y|>M]}).$$

其中 $M>0$，而 C_p 是仅同p有关的绝对常数。

3．预测

设有自变量X及因变量Y，X、Y可为多维。已知X有观察x，而Y的值尚未观察（或在观察X时尚不能观察Y）。要由x来预测Y的值。例如X为施肥量，Y为亩产量。当已知$X=x$时，Y的值要等到收获时才能观察。但人们希望在收获前，能从x预测Y将取何值。令 $L(y,a)$ 表示当Y实际取值为y，而预测为a时的损失。通常 L 取平方损失或绝对值损失这两种形式。设$\delta(x)$为某个预预规则，即当X取值x时用 $\delta(x)$ 预测Y之值。由于X取x有随机性，因而较为合理的是采用平均损失 $EL(Y,\delta(X))$作为预测规则 $\delta(\cdot)$ 的精度测度。称 $EL(Y,\delta(X))$ 为 δ 的（在 L 下）风险。若有规则 $\delta^*(\cdot)$，使得

$$EL(Y,\delta^*(X))=\inf_{\delta}EL(Y,\delta(X))\triangleq R^*, \tag{6·88}$$

则称 δ^* 为（在损失 L 下的）Bayes 预测，而称 R^* 为 Bayes 预测风险。上式的 inf 是取遍所有可能的预测规则。不难求得：当 $L(Y,a)=(Y-a)^2$ 时，$\delta^*(x)=E(Y|x)$；当 $L(Y,a)=|Y-a|$ 时，$\delta^*(x)=\xi\left(\frac{1}{2}|x\right)$。因而当$(X,Y)$的分布已知时，可以求得 Bayes 预测 δ^*。但在实际问题中(X,Y)的分布是未知的，只有来自(X,Y)的历史样本(X_i,Y_i)，$i=1,2,\cdots,n$. 例如(X_i,Y_i)是前几年的施肥量及亩产量。要求基于这组样本估计 $\delta^*(x)$。记任一这种估计为 $\delta_n^*(x)=\delta_n^*(x;X_1,Y_1,\cdots,X_n,Y_n)$。$\delta_n^*(x)$作为一个预测规则，显

然有 $EL(Y, \delta_n^*(X)) \geqslant R^*$. 直观上我们可以期望一个好的估计应使当样本容量不断增大时，其风险逐渐接近 R^*. 于是引出如下的定义。

定义6·8 设 L 为给定的损失，如一个估计 δ_n^* 使得

$$\lim_{n\to\infty} EL(Y, \delta_n^*(X)) = R^*, \qquad (6\cdot 89)$$

则称 δ_n^* 有（在 L 下）Bayes 相合性。又如 δ_n^* 是由权函数 $\{W_{ni}\}$ 所确定，则称 $\{W_{ni}\}$ 具有（在 L 下）Bayes 相合性。

下面我们考察 L 为平方损失及绝对值损失两种特殊情形。

当 $L(Y, a) = (Y-a)^2$，此时 Bayes 预测即给定 $X=x$ 时 Y 的条件期望。因而可由给定的权函数 $\{W_{ni}\}$ 构造如下的估计

$$\delta_{n1}^*(x) \triangleq \sum_{i=1}^{n} W_{ni}(x) Y_i. \qquad (6\cdot 90)$$

当 $L(Y, a) = |Y-a|$ 时，其相应的 Bayes 预测为给定 $X=x$ 时 Y 的条件中位数，可基于给定的 $\{W_{ni}\}$ 的定义估计

$$\delta_{n2}^*(x) \triangleq \xi_n\left(\frac{1}{2}\middle| x\right) = \sum_{i=1}^{n} W_{ni}(x) I_{[Y_i<\cdot]} \text{的中位数}, \qquad (6\cdot 91)$$

由定理 6·12 立即可得如下的大样本性质。

定理6·13 设 $\{W_{ni}\}$ 为矩相合的概率权函数，且

$$E|Y| < \infty,$$

若 $\xi\left(\frac{1}{2}|X\right)$ 以概率 1 唯一，则在绝对值损失下 $\{W_{ni}\}$ 有 Bayes 相合性（因而 δ_{n2}^* 有 Bayes 相合性）。

关于估计 δ_{n1}^* 也有类似的性质，我们有

定理6·14 设 $\{W_{ni}\}$ 为矩相合的概率权函数，且

$$E|Y|^2 < \infty,$$

则在平方损失下 $\{W_{ni}\}$ 为 Bayes 相合（因而估计 δ_{n1}^* 有 Bayes 相合性）。

证 因 $E|Y|^2 < \infty$，由矩相合的定义知

$$\lim_{n\to\infty} E[\delta_{n1}^*(X) - E(Y|X)]^2 = 0. \qquad (6\cdot 92)$$

注意到

$$
\begin{aligned}
EL(Y,\ \delta_{n1}^*(X)) &= E[Y-\delta_{n1}^*(X)]^2 \\
&= E[Y-E(Y|X)]^2 \\
&\quad + E[\delta_{n1}^*(X)-E(Y|X)]^2 \\
&\quad + 2E[Y-E(Y|X)] \\
&\quad \cdot[\delta_{n1}^*(X)-E(Y|X)] \\
&\triangleq J_{n1}+J_{n2}+J_{n3},
\end{aligned}
$$

又由（6·92），可得 $\lim\limits_{n\to\infty} J_{n2}=0$, $\lim\limits_{n\to\infty} J_{n3}=0$.

而 $E(Y|x)$ 是平方损失下的 Bayes 预测，故有

$$\lim_{n\to\infty} J_{n1}=R^*,$$

于是 $\lim\limits_{n\to\infty} EL(Y,\ \delta_{n1}^*(X))=R^*$. 定理证毕.

最后要指出的是，权函数方法有广泛的应用领域，以上涉及的三个应用专题只是其中的一部分，例如这种方法在非参数判别中也有重要应用，我们将在下一节中介绍。

§6·4　非参数判别

一、问题的提法

先看一些例子.

例6·1　某地区流行肝炎，为诊断某人有无肝炎，须抽血样进行化验，其化验结果（一般有若干项指标）可用一个向量X表示。用$Y=0$表示某人无肝炎，$Y=1$表示有肝炎。Y是一个取二值的类指标变量. 因而从该地区随机地抽取的一个个体，对应着随机向量（X，Y）的一个值. 显然化验结果X对判断有无肝炎（即Y的值）有很大作用. 但医学常识告诉我们，Y取何值尚不能据X所完全确定. 这受到医生的临床经验、化验手段是否精确可靠以及病人有无其它疾病等因素的影响. 因而这也是一个统计推断问题. 如果有一批历史资料$\{(X_i,\ Y_i),\ i=1,\ \cdots,\ n\}$可

用——这意味着以往曾对n个人作化验（确定X_i），并最终观察了每个人是否患肝炎，则对当前来接受化验的人，可按其化验结果X，参考已知样本$\{(X_i, Y_i), i=1, 2, \cdots, n\}$对其相应的$Y$值作出判别. 以此之故，$\{(X_i, Y_i), i=1, \cdots, n\}$常称为训练样本，意指它"训练了"人们如何去进行判断.

例6·2 设某种作物共有M个类，为对地球上该种作物进行大面积分类，通过卫星观察得遥感卫星照片数据. 每一张照片都对应有一个四维数据X，表示照片所在地区对四个光谱带的反射强度，而反射强度的大小与照片的色彩有关. 用Y表示某地区这种作物所属的类，则Y为取M值的类指标变量（例如Y取1, 2, $\cdots$, M）。在具体分类以前，需抽取其中的少数照片（例如n张），基于实地考察以分别确定这n张照片所涉及的地区实际上是属于哪一类，这样得到n个样品的观察指标$(X_i, Y_i), i=1, 2, \cdots, n$. 然后据此对剩下的照片进行逐张判别.

例6·3 为对某地区未经勘探的井位依有油、无油分类（用Y表示某井位的类指标），使用地震勘探技术可获得每个井位的地质数据（用一个多维指标向量X表示）. 在分类前选取少数的井位（例如n个）进行实地钻探，分别得到n个类指标$Y_1, \cdots, Y_n$. 由于钻井费用的昂贵，对别的井位在确定是否需要布井前，先要进行井位分类，因此要求对每一个井位的地质资料X，据已获得的样本$(X_i, Y_i), i=1, 2, \cdots, n$判定相应的$Y$值.

以上这些例子有这样几个共同特点：首先都有一个取有限值的类指标Y以及反映试验结果的指标向量X. Y同X有关，但知道X还不足以完全确定Y取值. 如在例6·3中知道某井位的地质资料X，还不足以判定该井位有油还是无油. 这是因为地质构造是非常复杂的，现代科学技术虽能分析大部分地质构造带变异的因素，但尚不能说已掌握所有的因素；其次某井位有油无油除了地质构造这一主要因素外，还受其它因素的影响，因而在X中并未包含Y的所有信息. 所以由X判定Y是一个统计问题. 其二

是，（X，Y）的联合分布是未知的，如在例6·1中，已知某人的血样为x，对其为肝炎的发病率究竟有多大，现代医学尚无确切的定量描述；其三是，由于（X，Y）的分布未知，在作判别前都先要有一组经过明确判定的样品。第i个样品的试验指标为X_i，类指标为Y_i。这里的Y_i是通过其它的试验手段得到的，如在例6·1中是通过临床观察，在例6·2是通过实地考察，而在例6·3则是通过钻井得到。它们构成日后进行判别工作的一个重要依据。最后，共同的问题是要求对新样品判定其所属类，即对任一给定的X，判定相应的Y值。这种例子还可举很多，文献上称这类问题为判别。由于对（X，Y）的分布类型并无特殊假定，问题属非参数性质，故可称为非参数判别。我们可以将判别问题一般化为：设某种对象可以划归为M（$\geqslant 2$）个类中的一个，而且只能一个。用Y表示类指标，而X为该对象的若干特征的指标变量。（X，Y）的分布未知，假定在以前鉴定过n个样品X_1，…，X_n，且分别知其所属的类为Y_1，…，Y_n。称

$$Z_n \triangleq \{(X_i, Y_i), i=1, \cdots, n\}$$

为训练样本。它是（X，Y）的独立同分布观察值。今有新样品并已测出其X指标，要利用X之值并借助于Z_n去判别此样品所属的类Y。

在上面的提法中，虽然也可将Y看作因变量，X看作自变量，但与预测问题不同的是：因变量Y只取有限个值，而且判别结果必须是这有限个可能值中的一个。而在预测问题中，因变量可取连续值，预测结果也可能越出Y取值范围之外。其次，类变量Y的各个可能值只是所代表类的标记，其值的大小并无意义。而预测问题中，Y一般是预测对象的实在值，其大小有通常的意义。因而在判别问题中，当Y实际属于i类，而判定为$i+1$类并不一定比判定为$i+2$类来得好。但当Y实际上为y，而预测其为$y+1$当然要比预测为$y+2$来得好。

我们在§6·1也曾提到过判别问题，依这里的一般提法可以看

成是一种特殊情况，即所属类的总数 $M=2$，且假定在给定 $Y=i$ 时，X 有条件密度。此时使用在那里提供的似然方法是有效的，在下面几段我们将分别介绍在一般情况下，非参数判别的有关概念及方法。

二、一般概念

往后总设 Y 为类变量，X 为指标变量，Y 只取 1，2…，M 为其可能值，X 为 d 维随机向量。判别问题的基本假定是：

(1) 给定 $Y=i$，X 有条件分布

$$F_i(x)=P(X\leqslant x|Y=i),\ i=1,\ 2,\ \cdots,\ M;$$

(2) Y 的分布为

$$p_i=P(Y=i),\ i=1,\ 2,\ \cdots,\ M;$$

(3) $Z_n\triangleq\{(X_i,\ Y_i),\ i=1,\ 2,\ \cdots,\ n\}$ 是来自 $(X,\ Y)$ 的独立同分布样本。

如假定给定 $Y=i$，X 有已知的条件密度 $f_i(x)$，$i=1$，…，M，则无须假定(2)及(3)，甚至 Y 不必看成是随机的。此时可采用§6·1所述的似然方法（对于一般的 M 如何处理放在后面介绍）判定 X 所属的类，当诸 $f_i(x)$ 为未知时，可以使用密度估计的方法估计之。其次，在通常的情况下，接受检验和判定的样品，其取得或来源是随机的，故假定 Y 为随机变量是合理的。在例 6·1 中，Y 取 0 或 1 是随机的；例 6·2 中，在实地考察前，某个地区的照片恰好属于哪一个类是随机的；在例 6·3 中，在钻井时某个井位有油、无油也是随机的。基于这一原因，假定(2)中的 $\{p_i\}$ 起着先验分布的作用，往后我们就称它为先验分布。$\{p_i\}$ 一般是基于过去的经验、一定的理论分析及专业知识得到的，而并不依赖试验结果 X 的信息。如例 6·1 可基于该地区在历史上曾经流行肝炎的资料得到肝炎发病率的估计，而同化验等试验手段无关。先验分布对判别结果有直接影响。因为若该地区发病率很低，则必须有更多的证据（相对于高发地区而言），才能更放心地判定某人有肝炎。再次，由假定(1)及(2)，$(X,\ Y)$ 的

联合分布也就随之确定，因而假定（3）应理解为 Z_n 是从此联合分布抽取的简单样本。由于（X，Y）的分布未知，知道X还不足以判定Y，样本 Z_n 是多次从 X_i 判定 Y_i 的实际经验总结，包含了（X，Y）联合分布的有关信息，因此对 Z_n 冠之以"训练样本"的称谓。训练样本在非参数判别中的作用是十分重要的，要求训练样本是独立同分布，这对实际选取训练样本施加了一定的限制。例如，不能把来源条件不同的样本混在一起。如在例6·2中，我们必须随机地抽取其中的几张照片进行实地考察，而不能贪图方便只选相邻近几个地区的照片，后者显然缺乏代表性。

在作了上述假定之后，统计模型就随之确定，随后最重要的是确定判别规则以及讨论判别规则的优良性。为此先要明确什么是判别规则？我们称一个定义在样本空间R^d 上的取值于{1，2，…，M}的函数 $g(\cdot)$为判别函数（或称判别规则），使得一旦新样本X有了一个给定值 x，就依此规则将它判归类 $g(x)$。依此定义，$g_0(x)=1$ 也是一种判别规则，但显然 g_0 是一种不足取的判别规则。然而如何辨别一个规则的好坏呢？这可以通过因错判而带来的损失去考察。我们用 $L(i, j)$表示当实际 $Y=i$，而判定j时所受的损失，下面就是一个最简单的，也是最常用的损失函数

$$L(i, j)=\begin{cases}0, & 当\ j=i;\\ 1, & 当\ j\neq i,\end{cases} \tag{6·93}$$

即判对时无损失，而只要是错判，其损失都一样，称此损失函数为0-1损失。我们约定，在本节的后文叙述中都使用这种0-1损失。对于给定的一个判别规则 g，由于Y及 $g(X)$都是随机变量，因而对使用g所蒙受的损失尚须取平均。称 $R(g)\triangleq EL(Y, g(X))$为规则$g$的风险。$R(g)$表示多次重复使用$g$作判别，所可能招致的平均损失。易见：

$$\begin{aligned}R(g)&=E(L(Y, g(X)))=P(L(Y, g(X))=1)\\ &=P(g(X)\neq Y),\end{aligned} \tag{6·94}$$

因而一个判别规则 g 的风险即是其错判概率。直观上，错判概率

是一个判别规则的有效程度的一种度量，错判概率越小越有效．往后我们还采用另一种度量，即后验风险或后验错判概率：

$$r(g;\ x) \triangleq P(g(X) \neq Y | X = x), \tag{6·95}$$

后验错判概率 $r(g;\ x)$ 的含义是，在具同一指标 x 的样品群中，用判别规则 g 时的错判概率。$R(g)$ 与 $r(g;\ x)$ 之间有如下关系：

$$R(g) = \int r(g;\ x)\, dF(x), \tag{6·96}$$

其中 F 是 X 的分布函数。由上述可知，在 0-1 损失下，错判概率愈小，该判别规则愈佳。这引致下面的定义。

定义6·9 设 g^* 是一个给定的判别规则，如对任意一个判别规则 g，都有

$$P(g^*(X) \neq Y) \leqslant P(g(X) \neq Y) \tag{6·97}$$

则称 g^* 为最佳判别或称 Bayes 判别规则，而称

$$R^* \triangleq P(g^*(X) \neq Y) \tag{6·98}$$

为 Bayes 风险或 Bayes 错判概率。

其所以把最佳判别称为 Bayes 判别，是因为上文指出的一个事实，即 Y 的分布带有先验分布的性质。在一般情况下，由于 $(X,\ Y)$ 的联合分布不知道，Bayes 判别规则无法求出。若已知 Y 的分布为 $p_i = P(Y = i), i = 1, \cdots, M$，以及给定 $Y = i$ 时 X 的条件密度 $f_i(x)$，则不难求得：在已知 $X = x$ 时，Y 的条件分布为

$$\begin{aligned}\eta_i(x) &\triangleq P(Y = i | X = x)\\ &= p_i f_i(x) \Big/ \sum_{j=1}^{M} p_j f_j(x),\quad i = 1, \cdots, M,\quad x \in R^d\end{aligned} \tag{6·99}$$

而 Bayes 判别规则 g^* 为

$$g^*(x) = i,\ 若\ \eta_i(x) = \max_{1 \leqslant j \leqslant M} \eta_j(x). \tag{6·100}$$

换言之，那个类的后验（条件）概率最大，即将样品判归该类．这规则在直观上显然。我们约定，当使 $\eta_i(x)$ 达到最大的 i 不

止一个时，取标号最小的那一个。(6·100)的确定是基于 Bayes 统计的基本原则，即使后验风险最小原则。此处易直接证明：g^*确为错判概率最小之判别规则。若(X, Y)的分布未知(这时$\{p_i\}$或$\{f_i\}$未知)，无法根据(6·100)确定 Bayes 规则g^*，一种自然的途径是使用训练样本估计后验分布$\{\eta_i(x)\}$。以$\{\eta_{ni}(x)\}$记这样一个估计，用η_{ni}代替η_i于(6·100)，而得出如下的判别规则：

$$g_n^*(x)=g_n^*(x; Z_n)=i, \text{ 如 } \eta_{ni}(x)=\max_{1\leq j\leq M}\eta_{nj}(x), \tag{6·101}$$

显然有一种估计$\{\eta_{ni}(x)\}$就有一个规则$g_n^*(x)$。由于此法也是基于后验分布，故称此种构造判别规则的方法为 Bayes 方法(Bayes 方法指构造判别规则的一类方法，不要与作为最佳判别法的 Bayes 判别规则混为一谈)。往后还要介绍另一种基于Z_n构造判别规则的方法，即最近邻法。Bayes 法与最近邻法构成非参数判别的两种主要方法。

现设$g_n(x)\triangleq g_n(x; Z_n)$为基于$Z_n$的任意一个判别规则，即对固定的$Z_n$，$g_n(x)$作为$x$的函数是一个判别函数，因而依 Bayes 风险的定义，应有

$$P(g_n(X)\neq Y)\geq R^*, \tag{6·102}$$

一般说来等号不成立。但有理由期望：一个好的规则g_n，当训练样本大小n不断增加时，其错判概率可以随之逼近R^*。这是因为一旦有了一个"无穷大"的样本(X_1, Y_1)，(X_2, Y_2)，…，则相当于对(X, Y)的分布有了完全的知识，即知道了(X, Y)的分布，因而也就可确切地定出 Bayes 判别。我们也可以从另一角度评价g_n的优劣。记

$$L_n\triangleq L_n(g_n, Z_n)=P(g_n(X)\neq Y|Z_n), \tag{6·103}$$

$$r_n(x)\triangleq r_n(g_n, x)\triangleq P(g_n(X)\neq Y|X=x), \tag{6·104}$$

$$r^*(x)\triangleq r^*(g^*, x)\triangleq P(g^*(X)\neq Y|X=x), \tag{6·105}$$

其中g^*为 Bayes 规则。L_n、$r_n(x)$、$r^*(x)$都是不同条件下的

后验错判概率，则有

$$R^* = Er^*(X),$$
$$R(g_n) = EL_n(g_n, Z_n) = Er_n(g_n, X).$$

而且由（6·105）知

$$\begin{aligned} r^*(x) &= 1 - P(g^*(X) = Y \mid X = x) \\ &= 1 - P(Y = g^*(x) \mid X = x) \\ &= 1 - \eta_{g^*(x)}(x) \\ &= 1 - \max_{1 \leqslant i \leqslant M} \eta_i(x), \end{aligned} \qquad (6\text{·}106)$$

因此，例如当 $n \to \infty$ 时，有

$$r_n(g_n, X) \xrightarrow[(\text{或a.s.})]{p} r^*(X),$$

或者

$$L_n(g_n, Z_n) \xrightarrow[(\text{或a.s.})]{p} R^*.$$

可以认为规则 g_n 在某种程度上是好的。量 L_n 及 $r_n(x)$ 都有其实际意义，特别是 L_n。L_n 表示在给定训练样本 Z_n 的条件下，反复使用规则 g_n（即多次使用同一个 Z_n）作判别，其条件平均损失。在许多实际问题中，出于种种原因，同一训练样本要反复使用多次。如例 6·3 中，Z_n 是通过实地钻井得到，钻每口井耗资甚大，因而对 Z_n 甚为珍惜。L_n 作为一种有效程度的度量恰好适应这一要求。在这种情况下，作为无条件平均 $R(g_n)$ 显得没有什么实用价值。

三、Bayes 方法

上一段已经提到，Bayes 方法的要旨是：基于训练样本 Z_n 估计后验分布 $\{\eta_i(x)\}$，然后依（6·101）构造判别规则，使错判概率尽可能地小。下面介绍两种估计后验分布的方法，即权函数法及概率密度估计法。

1．权函数法

此法将估计 $\{\eta_i(x)\}$ 的问题化为回归估计。我们引进新变量

$$\tilde{Y}=\begin{cases}1, & 当 Y=i; \\ 0, & 当 Y\neq i,\end{cases}$$

则

$$\eta_i(x)=P(Y=i|X=x)=E(\tilde{Y}|X=x).$$

对给定的权函数

$$W_{nj}(x)=W_{nj}(x;\ X_1,\ \cdots,\ X_n),\ j=1,\ 2,\ \cdots,\ n,$$

定义 $\eta_i(x)$ 的估计为

$$\eta_{ni}(x)=\sum_{j=1}^{n}W_{nj}(x)I_{(Y_j=i)}, \tag{6·107}$$

然后据(6·101)确定 g_n^*。由此构造方法可知：每一个规则 g_n^* 由权函数 $\{W_{ni}\}$ 所完全确定，下面是这一方法的基本性质。

定理6·15 不论对什么样的权函数，由上述方法构造的规则 $g_n^*(x)=g_n^*(x;\ Z_n)$，都有

$$0\leqslant r_n(g_n^*,\ x,\ Z_n)-r^*(x)$$
$$\leqslant 2\sum_{i=1}^{M}|\eta_{ni}(x)-\eta_i(x)|, \tag{6·108}$$

其中

$$r_n(g_n^*,\ x,\ Z_n)=p(g_n^*(X)\neq Y|X=x,\ Z_n), \tag{6·109}$$

证明 (6·108)的左边不等式容易，因当固定 Z_n 后，$g_n^*(x;\ Z_n)$ 是关于变元 x 的判别函数，再给定 $X=x$ 时的后验错判概率即为 $r_n(g_n^*,\ x,\ Z_n)$，当然不能小于给定 $X=x$ 时的 Bayes 判别 g^* 的后验错判概率。为证右边不等式，注意由$(X,\ Y)$与 Z_n 独立及 g_n^* 的构造，有

$$r_n(g_n^*,\ x,\ Z_n)=1-\eta_{n,g_n^*(x,Z_n)}(x),$$

再由(6·106)式知

$$r_n(g_n^*,\ x,\ Z_n)-r^*(x)$$
$$=\max_{1\leqslant j\leqslant M}\eta_j(x)-\eta_{n,g_n^*(x,Z_n)}(x).$$

下面我们注意一个初等不等式：设 $a_1,\ \cdots,\ a_M$ 及 $b_1,\ \cdots,\ b_M$ 为 $2M$ 个常数，且 $a_i=\max\limits_{1\leqslant j\leqslant M}a_j$，$b_k=\max\limits_{1\leqslant j\leqslant M}b_j$，则

$$|a_i - b_k| \leqslant \sum_{j=1}^{M} |a_j - b_j|. \tag{6·110}$$

在此不等式中令 $a_i = \eta_i(x)$，$b_i = \eta_{ni}(x)$，注意到

$$\eta_{n, g^*_{n(x, Z_n)}}(x) = \max_{1 \leqslant j \leqslant M} \eta_{nj}(x),$$

于是由（6·110）得到

$$r_n(g_n, x, Z_n) - r^*(x) = \left|\max_{1 \leqslant j \leqslant M} \eta_j(x) - \eta_{n g^*_{n(x, Z_n)}}(x)\right|$$

$$\leqslant \sum_{j=1}^{M} |\eta_j(x) - \eta_{n,j}(x)|. \tag{6·111}$$

此即右端不等式成立，定理证毕。

由此定理可得下述推论。

推论 若$\{W_{ni}\}$是矩相合，则

$$\lim_{n\to\infty} R(g_n^*) = R^*.$$

证明 由矩相合性及 η_{ni} 的构造（6·107）知

$$\lim_{n\to\infty} E|\eta_{ni}(X) - \eta_i(X)| = 0, \quad i = 1, \cdots, M,$$

因而由（6·111）知

$$\lim_{n\to\infty} E|r_n(g_n^*; X, Z_n) - r^*(X)| = 0,$$

但由此即得

$$\lim_{n\to\infty} R(g_n^*) = \lim_{n\to\infty} Er_n(g_n^*; X, Z_n)$$

$$= Er^*(X) = R^*.$$

此推论表明，由矩相合权函数构造的判别规则有渐近 Bayes 性。

2。密度估计法

此法假定当给定 $Y=i$ 时，X有条件密度 $f_i(x)$，$i=1, 2, \cdots, M$，由 Bayes 公式可知给定$X=x$时，Y有后验分布

$$\eta_i(x) \triangleq P(Y=i|X=x)$$

$$= p_i f_i(x) \Big/ \sum_{j=1}^{M} p_j f_j(x), \quad i=1, \cdots, M, \tag{6·112}$$

为简单计，考虑 $M=2$ 的情形。令

$$D(x) = p_1 f_1(x) - p_2 f_2(x), \tag{6·113}$$

此时 Bayes 规则 g^* 可表为

$$g^*(x)=\begin{cases}1, & \text{当 } D(x)\geqslant 0;\\ 2, & \text{当 } D(x)<0,\end{cases} \tag{6·114}$$

如能基于 Z_n 估计 $D(x)$，记任一这种估计为

$$D_n(x)\triangleq D_n(x;\ Z_n),$$

再仿照（6·114）定义判别规则，

$$g_n^*(x)=\begin{cases}1, & \text{当 } D_n(x)\geqslant 0;\\ 2, & \text{当 } D_n(x)<0,\end{cases} \tag{6·115}$$

显然有一个 $D(x)$ 的估计 $D_n(x)$，就有一个上述的判别规则 g_n^*，因而 g_n^* 由 $D_n(x)$ 所完全确定。Bayes 规则 g^* 的 Bayes 风险为

$$\begin{aligned}R^* &= P(g^*(X)\neq Y)\\ &= P(Y=1,\ D(X)<0)+P(Y=2,\ D(X)\geqslant 0)\\ &= p_1P(D(X)<0\,|\,Y=1)+p_2P(D(X)\geqslant 0\,|\,Y=2)\\ &= p_1\int_H f_1(x)\,dx+p_2\int_{H^c} f_2(x)\,dx \qquad (6·116)\\ &= p_1-\int_{H^c} D(x)\,dx, \qquad (6·117)\end{aligned}$$

其中

$$H=\{x: D(x)<0\}=\{x: p_1f_1(x)<p_2f_2(x)\} \tag{6·118}$$

而规则 g_n^* 在固定 Z_n 时的后验错判概率为

$$\begin{aligned}L_n\triangleq L_n(g_n^*,\ Z_n) &= P(g_n^*(X)\neq Y\,|\,Z_n)\\ &= p_1\int_{H_n} f_1(x)\,dx+p_2\int_{H_n^c} f_2(x)\,dx \qquad (6·119)\\ &= p_1-\int_{H_n^c} D(x)\,dx, \qquad (6·120)\end{aligned}$$

其中

$$H_n=\{x: D_n(x)<0\}. \tag{6·121}$$

由（6·116）及（6·119）式，即有

$$R^*=\int \min(p_1f_1(x),\ p_2f_2(x))\,dx$$

$$=\int_{H_n}+\int_{H_n^c}\min(p_1f_1(x),\ p_2f_2(x))dx$$
$$\leqslant p_1\int_{H^n}f_1(x)dx+p_2\int_{H_n^c}f_2(x)dx$$
$$=L_n. \qquad (6\cdot122)$$

上式表明：不论估计 $D_n(x)$ 是怎样得到的，由（6·115）所确定的规则 g_n^*，其后验错判概率总是不低于 R^*，这是与 Bayes 风险的含义一致的。而且我们还有

$$0\leqslant L_n-R^*$$
$$=\int D(x)(I_{H^c}(x)-I_{H_n^c}(x))dx. \qquad (6\cdot123)$$

规则 g_n^* 有如下的大样本性质。

定理6·16 不论 $D_n(x)$ 是用什么方法得到的估计，如果当 $n\to\infty$

$$\int(D_n(x)-D(x))^2dx\xrightarrow{P}0\ (\text{或a.s.}), \qquad (6\cdot124)$$

则

$$L_n\xrightarrow{P}R^*\ (\text{或a.s.}). \qquad (6\cdot125)$$

证明 记X的边缘密度为 $f(x)$，则

$$f(x)=p_1f_1(x)+p_2f_2(x)$$

对任一d维 Borel 集 A，用$|A|$表示 A 的体积。对任给 $\varepsilon>0$，取 d 维有界 Borel 集B充分大，使得

$$\int_B f(x)dx>1-\varepsilon/2. \qquad (6\cdot126)$$

因为

$$D_n(x)\geqslant0\Rightarrow I_{H_n}(x)=0$$
$$\Rightarrow I_{H^c}(x)-I_{H_n^c}(x)\leqslant0;$$
$$D_n(x)<0\Rightarrow I_{H_n}(x)=1$$
$$\Rightarrow I_{H^c}(x)-I_{H_n^c}(x)\geqslant0,$$

故而

$$-D_n(x)(I_{H^c}(x)-I_{H_n^c}(x))\geqslant0,$$

于是由（6·123）即知

$$
\begin{aligned}
0 &\leqslant L_n - R^* \\
&\leqslant \int_B (D(x) - D_n(x))(I_{B^c}(x) - I_{B_n^c}(x))dx \\
&\quad + \int_{B^c} D(x)(I_{B^c}(x) - I_{B_n^c}(x))dx \\
&\triangleq J_{n1} + J_{n2}.
\end{aligned}
\qquad (6\cdot127)
$$

而 $|J_{n1}| \leqslant \int_B |D(x) - D_n(x)| dx$

$$
\leqslant \left[\int_B |D(x) - D_n(x)|^2 dx\right]^{1/2} \sqrt{|B|},
$$

又由

$$
\begin{aligned}
|J_{n2}| &\leqslant \int_{B^c} |D(x)| dx \\
&\leqslant \int_{B^c} f(x) dx < \varepsilon/2,
\end{aligned}
$$

因而

$$
\begin{aligned}
0 &\leqslant L_n - R^* \\
&\leqslant \left[\int_B |D_n(x) - D(x)|^2 dx\right]^{1/2} \sqrt{|B|} + \varepsilon/2.
\end{aligned}
$$

由假设条件（6·124），先令 $n\to\infty$，再让 $\varepsilon\to0$ 即得证(6·125)，定理证毕。

使用定理6·16，需验证条件（6·124），这取决于 $D_n(x)$ 的结构及原模型的条件。

现讨论 $D(x)$ 的估计，一种自然的方法是基于 Z_n 分别估计 p_1、p_2 及 f_1、f_2。记 $l_n = {}^\#\{i: 1\leqslant i\leqslant n, Y_i=1\}$，用 $\frac{l_n}{n}$ 估计 p_1，$\frac{n-l_n}{n}$ 估计 p_2，再令 K 为核函数，选取窗宽 $h_n>0$。记 $\Lambda_1=\{i:1\leqslant i\leqslant n, \ Y_i=1\}$，$\Lambda_2=\{i:1\leqslant i\leqslant n, Y_i=2\}$，则可分别定义 f_1、f_2 的核估计为

$$f_{n1}(x)=\frac{1}{l_n h_n^d}\sum_{i=1}^{n} I_{[Y_i=1]} K\left(\frac{x-X_i}{h_n}\right)$$

$$=\frac{1}{l_n h_n^d}\sum_{i\in\Lambda_1} K\left(\frac{x-X_i}{h_n}\right),$$

$$f_{n2}(x)=\frac{1}{(n-l_n)h_n^d}\sum_{i=1}^{n} I_{(Y_i=2)} K\left(\frac{x-X_i}{h_n}\right)$$

$$=\frac{1}{(n-l_n)h_n^d}\sum_{i\in\Lambda_2} K\left(\frac{x-X_i}{h_n}\right).$$

然后依 $D(x)$ 的构造，定义 $D(x)$ 的估计为

$$D_n(x)=\frac{l_n}{n}\frac{1}{l_n h_n^d}\sum_{i\in\Lambda_1} K\left(\frac{x-X_i}{h_n}\right)$$

$$-\frac{n-l_n}{n}\frac{1}{(n-l_n)h_n^d}\sum_{i\in\Lambda_2} K\left(\frac{x-X_i}{h_n}\right)$$

$$=\frac{1}{nh_n^d}\left(\sum_{i\in\Lambda_1} K\left(\frac{x-X_i}{h_n}\right)\right.$$

$$\left.-\sum_{i\in\Lambda_2} K\left(\frac{x-X_i}{h_n}\right)\right), \tag{6·128}$$

对此 $D_n(x)$，我们有

定理6·17 设 f_1、f_2 一致连续，而 h_n 满足

$$h_n\to 0,\ nh_n^d\to\infty,\ 当\ n\to\infty.$$

则

$$\int\left[D_n(x)-D(x)\right]^2 dx \xrightarrow{p} 0,\ 当\ n\to\infty,$$

因而

$$L_n \xrightarrow{P} R^*,\ 当 n\to\infty.$$

其证明方法类似于引理 6·3 及定理 6·2，这里从略。

若对 $D(x)$ 变形还可以得到别的估计方法。但其思想都是分别用密度估计方法估计条件密度或边际密度，用频率估计先验概率，而且在适当的条件下，也可以证明类似于定理 6·17 的结果，故在此不作一一介绍。下面举一个实例。

例6·4（例 6·3 续） 地质数据是一个三维变量 $X=(X^{(1)},$

$X^{(2)}$, $X^{(3)}$)，其中$X^{(1)}$是切向运动指标经过离散处理后只取1、2、3三个值；$X^{(2)}$是反映有无断层的指标，只取0、1二个值；而$X^{(3)}$是表征有无陡带的指标，也是0-1变量，其样本空间是R^3中一个有限子集(只含12个点子)。设Y是类变量，$Y=0$表示无油，而$Y=1$表示有油，因而$M=2$。记$p_i=P(Y=i)$，$i=0, 1$，$f_0(x)=P(X=x|Y=0)$，$f_1(x)=P(X=x|Y=1)$，而

$$D(x)=p_1f_1(x)-p_0f_0(x), \tag{6·129}$$

此时 Bayes 规则为

当$D(x)\geqslant 0$，判x为有油，

当$D(x)<0$，判x为无油。

今采用邻近已开钻地区的资料，得到$n=63$的训练样本。注意到$p_0f_0(x)=P(X=x, Y=0)$，$p_1f_1(x)=P(X=x, Y=1)$，因而可直接用频率估计概率$P(X=x, Y=0)$及$P(X=x, Y=1)$，而无须分别估计p_0、p_1、f_0、f_1。记如此得到的$D(x)$的估计为$D_n(x)$，其判别规则为

当$D_n(x)\geqslant 0$，判x为有油，

当$D_n(x)<0$，判x为无油。

因样本空间只含12个点，$M=2$，可对每个x计算$D_n(x)$，其计算程序并不复杂。然后对12种地质类型依有油，无油分类，造出一张分类表。对每一新井位，依测得的地质数据x，从分类表中可找到其类别。依强大数律，对每一给定x，有

$$D_n(x)\longrightarrow D(x), \text{ a.s. 当 } n\to\infty,$$

$$\eta_{ni}(x)\longrightarrow\eta_i(x), \text{ a.s. 当 } n\to\infty, \ i=0, 1,$$

$$f_n(x)\to f(x), \text{ a.s. 当 } n\to\infty.$$

其中$\eta_i(x)=P(Y=i|X=x)$，$i=0, 1$，$f(x)=p_0f_0(x)+p_1f_1(x)$，而η_{ni}，f_n分别是其频率估计。易知

$$0\leqslant L_n-R^*\leqslant 2\sum_{i=0}^{1}\sum_{x\in\mathscr{X}}|\eta_{ni}(x)-\eta_i(x)|f(x), \tag{6·130}$$

因而（由样本空间 $\mathscr{X}$ 有限）$L_n \longrightarrow R^*$, a.s., 当 $n\to\infty$. 依我们的数据，对 L_{63} 进行估计．算得 L_{63} 近似为 0.115。相对于63个训练样本来说，这个值已比较理想．再使用已得的分类表，用邻近地区已得数据加以印证，其错判的频率大约为 0.201。之所以出现这个情况，一者频率本身同概率总有差异；二是在计算中，对某些 x，$D_n(x)$ 的绝对值很接近 0，很难判定其所属类．因而有较大的误差，但即使这样，同不使用此法而单凭经验公式相比有较大的改善。

四、近邻判别

近邻判别法最早是由 Fix 和 Hodges 在 1951-1952 年引进的。其基本出发点是，对于给定的 $X=x$ 及训练样本 Z_n，在判别其所属类时，只有最接近 x 的那些样本才起作用。我们可将其方法要旨表述如下：在 R^d 中引进距离 $\rho(x, y)$，对给定的 $X=x$，按照 $\{\rho(x, X_i), i=1, \cdots, n\}$ 的上升次序将指标样本 $X_1, \cdots, X_n$ 重新排为 $X_{R1}, \cdots, X_{Rn}$，与之匹配的 $\{Y_i\}$ 记为 $Y_{R1}, \cdots, Y_{Rn}$.再定下一个自然数 k（$1\leqslant k<n$），在 $Y_{R1}, \cdots, Y_{Rk}$ 中用"多数选举"原则决定 x 所属的类，即当 $Y_{R1}, \cdots Y_{Rk}$ 中等于 i 的个数最多，判 x 为 i 类。我们将之确切地表示为下述的定义。

定义6·10 记 $l_i={}^*(\{j: Y_{R_j}=i, j=1,\cdots, k\})$, $i=1,\cdots, M$. 定义判别规则 $g_n^{(k)}(x) \triangleq g_n^{(k)}(x; Z_n)$ 为

$g_n^{(k)}(x)=i$，若 l_i 是 $l_1,\cdots, l_M$ 中唯一最大者。若同时有若干个达到最大，例如 $l_{i1}, \cdots, l_{ic}$，则依等概率在 $i_1, \cdots i_c$ 中随机决定一个为 $g_n^{(k)}(x)$．称如此定义的 $g_n^{(k)}$ 为 k-近邻判别规则（k-N.N.）。当 $k=1$ 时则简称近邻规则（N.N.），且记 $g_n^{(1)}$ 为 g_n。

依此定义，近邻规则是

$$当\ Y_{R1}=i\ 时，判\ x\ 为\ i\ 类. \qquad (6\cdot131)$$

在定义 6·10 中，要求依 $\{\rho(x, X_i)\}$ 的上升次序重新排序。

在出现"结"的时候，我们仍采用"足标靠前"的原则。下面就 $k=1$ 及一般的 k 这两种情形分别讨论有关概念及性质。

1. N.N. 法 沿用前面的记号，并记 $X_1, \cdots, X_n$ 中与 X 最接近者为 X'_n，与之匹配的 Y 记为 Y'_n。改记 N.N. 错判概率为

$$R_n \triangleq P(Y'_n \neq Y). \tag{6·132}$$

我们有如下性质。

定理6·18 设 $M=2$，$\eta_i(x)$ 连续，$i=1, 2$，则有

$$R^* \leqslant R \leqslant 2R^*(1-R^*), \tag{6·133}$$

其中

$$R = \lim_{n\to\infty} R_n. \tag{6·134}$$

在证明定理之前，先给出一个注解。因

$$P(g^*(X)=Y|X=x) = \max(\eta_1(x), \eta_2(x)) \geqslant \frac{1}{2},$$

有 $R^* = EP(g^*(X) \neq Y|X) \leqslant \frac{1}{2}$，于是

$$2R^*(1-R^*) \geqslant 2R^*\left(1-\frac{1}{2}\right) = R^*,$$

因此（6·133）式是合理的。

证明 先证明一个往后要多次用到的事实，即除去一个 F 零测集外的 x，有

$$X'_n(x) \longrightarrow x, \text{ a.s. 当} n\to\infty, \tag{6·135}$$

其中 F 为 X 的边缘分布，而 $X'_n(x)$ 为当 $X=x$ 时 X'_n 的相应标记。

事实上，若记 C 为 F 的支撑集，则 $F(C)=1$。任取 $x\in C$，对任给的 $\varepsilon>0$ 有

$$\begin{aligned} P(\rho(x, X'_n(x)) \geqslant \varepsilon) &= P(\rho(x, X_1) \geqslant \varepsilon, \cdots, \rho(x, X_n) \geqslant \varepsilon) \\ &= [P(\rho(x, X) \geqslant \varepsilon)]^n = [1-P(\rho(x, X) < \varepsilon)]^n \end{aligned}$$

因 $x\in C$，有 $P(\rho(x, X)<\varepsilon)>0$，从而

$$0 \leqslant r \triangleq 1-P(\rho(x, X)<\varepsilon) < 1.$$

于是 $\sum_{n=1}^{\infty} P(\rho(x, X'_n(x)) \geqslant \varepsilon) = \sum_{n=1}^{\infty} r^n < \infty$。由 Borel-Cantelli 引理即得证（6·135）式.

现回到定理的证明。易知

$$r_n(x \cdot x'_n) \triangleq P(Y'_n \neq Y | X = x, \ X'_n = x'_n)$$
$$= \eta_1(x)\eta_2(x'_n) + \eta_2(x)\eta_1(x'_n),$$

由 $\eta_i(x)$ 的连续性及（6·135）可知：对 a.e. $x[F]$，

$$\lim_{n\to\infty} r_n(x, X'_n) = 2\eta_1(x)\eta_2(x), \text{a.s.} \tag{6·136}$$

但由 $r^*(x) = \min(\eta_1(x), \ \eta_2(x)) \leqslant \frac{1}{2}$，可得

$$r^*(x) \leqslant 2\eta_1(x)\eta_2(x) = 2r^*(x)(1 - r^*(x))$$
$$\leqslant 2r^*(x). \tag{6·137}$$

再由（6·136）使用控制收敛定理得到

$$R \triangleq \lim_{n\to\infty} R_n = \lim_{n\to\infty} Er_n(X, \ X'_n) = E[2\eta_1(X)\eta_2(X)],$$

结合（6·137）式，即有

$$R^* = Er^*(X) \leqslant R = E[2\eta_1(X)\eta_2(X)]$$
$$= E[2r^*(X)(1 - r^*(X))]$$
$$\leqslant 2\{Er^*(X) - [Er^*(x)]^2\}$$
$$= 2R^*(1 - R^*).$$

定理证毕.

不等式（6·133）提供了这样一个有趣事实：当有了一个“无穷”样本时，本应采用 Bayes 规则，相应的错判概率为 R^*；如仍坚持用 N.N. 法，则错判概率至多为 $2R^*(1-R^*) \leqslant 2R^*$。这可以粗略地解释为：即使有极大数目的样本，但与$X$最接近的那一个，其信息量占到全样本的一半。这一解释为使用 N.N. 原则提供了重要依据：其次，从不等式（6·133）可知，$R^* = 0$，当且仅当 $R = 0$；$R^* = \frac{1}{2}$ 当且仅当 $R = \frac{1}{2}$。这就是说，在完全确定及完全不确定这两种极端情形下，从大样本角度看 N. N. 判别

与 Bayes 判别是相同的。

上述结果不难推广到一般的M。即在关于$\{\eta_i(x)\}$连续性的假定下，有

$$R^* \leqslant R \leqslant R^* \left(2 - \frac{M}{M-1} R^*\right) \tag{6·138}$$

例6·5 设样本空间为开区间(0, 1)，$M=2$，$p_1=p_2=\frac{1}{2}$，

$$f_1(x)=\begin{cases}2(1-x), & \text{当 } 0<x<1; \\ 0, & \text{其它的 } x,\end{cases}$$

$$f_2(x)=\begin{cases}2x, & \text{当 } 0<x<1; \\ 0, & \text{其它的 } x,\end{cases}$$

经简单计算可得 $R=\frac{1}{3}$，$R^*=\frac{1}{4}$。显然满足定理6·18的不等式(6·133)。

关于N.N.法的另一性质是：

定理6·19 设$M=2$，$f_i(x)$连续$i=1, 2$。则

$$L_n \triangleq P(Y_n' \neq Y \mid Z_n) \xrightarrow{P} R, \quad \text{当 } n\to\infty. \tag{6·139}$$

证明从略。

2．k-N.N.法　为使叙述简便起见，假设$M=2$。本段讨论k-N.N.规则的一些性质。为此先证明一个往后要用到的预备事实。

引理6·5 设事件$A_1, \cdots, A_k$相互独立，$p_i=P(A_i)\, i=1, \cdots, k$。对某个$p\in(0, 1)$，$1\leqslant i\leqslant k$，记

$$b(k;\ p,\ i)=\binom{k}{i}p^i(1-p)^{k-i} \tag{6·140}$$

$C(k;\ p_1, \cdots, p_k, i)=P(A_1, \cdots, A_k$恰好出现$i$个$)$，则对$\{1, 2, \cdots, k\}$的任一子集$A$有　(6·141)

$$\left|\sum_{i\in A} C(k;p_1, \cdots, p_k, i) - \sum_{i\in A} b(k;p, i)\right|$$

$$\leqslant \sum_{i=1}^{k} |p_i - p|. \tag{6·142}$$

证明 设$\xi_1, \cdots, \xi_k$独立同(0, 1)上均匀分布。令

$$\eta_i = I_{\{\xi_i < p_i\}},\quad \eta_i' = I_{\{\xi_i < p\}},\quad i=1,\ \cdots,\ k$$

则对任一$1 \leqslant j \leqslant k$,

$$P\left(\sum_{i=1}^{k} \eta_i = j\right) = C(k; p_1,\ \cdots,\ p_k,\ j),$$

$$P\left(\sum_{i=1}^{k} \eta_i' = j\right) = b(k; p,\ j).$$

因而

$$(6\cdot142)\text{式的左端} = \left| P\left(\sum_{i=1}^{k} \eta_i \in \wedge\right) - P\left(\sum_{i=1}^{k} \eta_i' \in \wedge\right) \right|$$

$$\leqslant P\left(\bigcup_{i=1}^{k} \{\eta_i \neq \eta_i'\}\right) \leqslant \sum_{i=1}^{k} P(\eta_i \neq \eta_i') = \sum_{i=1}^{k} |p_i - p|,$$

引理证毕。

重新记 $X_1' = X_{R_1}, \cdots, X_k' = X_{R_k}$, $Y_1' = Y_{R_1}, \cdots, Y_k' = Y_{R_k}$,
记

$$r_n(x; x_1',\ \cdots,\ x_k') = P(g_n^{(k)}(X) \neq Y \mid X = x,\ X_i' = x_i',\ i=1,\ \cdots,\ k),$$

$$A_i = \{Y_i' = 1\},\quad i = 1,\ 2,\ \cdots,\ k.$$

则在给定 $X = x$, $X_i' = x_i'$, $i=1, \cdots, k$ 时, $A_1, \cdots, A_k$ 条件独立, 且 $P(A_i | X_j' = x_j',\ j=1, \cdots, k) = \eta_1(x_i')$, $i=1, \cdots, k$.
依 $g_n^{(k)}$ 的定义, 有

$$P(g_n^{(k)}(x) = 1 \mid X = x,\ X_i' = x_i',\ i=1,\ \cdots,\ k)$$

$$= P\left(A_1, \cdots, A_k\text{ 至少出现}\frac{k}{2}+1\text{个} \mid X_j' = x_j', j=1\cdots, k\right)$$

$$+ P\left(A_1, \cdots, A_k\text{ 恰好出现}\frac{k}{2}\text{个} \mid X_j' = x_j',\ j=1,\ \cdots, k\right)\cdot\frac{1}{2}$$

$$= \sum_{\frac{k}{2} < i \leqslant k} C(k; \eta_1(x_1'), \cdots,\ \eta_1(x_k'),\ i)$$

$$+ \frac{1}{2} C\left(k; \eta_1(x_1'),\ \cdots,\ \eta_1(x_k'),\ \frac{k}{2}\right)$$

此处及下面都约定：当$\frac{k}{2}$非整数时

$$b\left(k;\ p,\ \frac{k}{2}\right)=0,\quad c\left(k;\ p_1,\ \cdots p_k,\ \frac{k}{2}\right)=0.$$

同理

$$P(g_n^{(k)}(x)=2|X=x,\ X_i'=x_i',\ i=1,\ \cdots,\ k)$$

$$=\sum_{0<i<\frac{k}{2}} C(k;\eta_1(x_1'),\ \cdots,\ \eta_1(x_k'),\ i)$$

$$+\frac{1}{2}C\left(k;\eta_1(x'),\ \cdots,\ \eta_1,\ (x_k'),\ \frac{k}{2}\right)$$

因而

$$r_n(x;x_1',\ \cdots,\ x_k')=P(g_n^{(k)}(x)=1,\ Y=2|X$$

$$=x,X_i'=x_i',\ i=1,\ \cdots,\ k)$$

$$+P(g_n^{(k)}(x)=2,\ Y=1|X=x,\ X_i'=x_i',\ i=1,\ \cdots,\ k)$$

$$=\eta_2(x)P(g_n^{(k)}(x)=1|X=x,\ X_j'=x_j',\ j=1,\ \cdots k)$$

$$+\eta_1(x)P(g_n^{(k)}(x)=2|X=x,\ X_j'=x_j',\ j=1,\cdots,k)$$

$$=\eta_2(x)\sum_{\frac{k}{2}<i<k} C(k;\eta_1(x_1'),\ \cdots,\ \eta_1(x_k'),\ i)$$

$$+\eta_1(x)\sum_{0<i<\frac{k}{2}} C(k;\eta_1(x_1'),\ \cdots,\ \eta_1(x_k'),\ i)$$

$$+\frac{1}{2}C\left(k;\eta_1(x_i'),\ \cdots,\ \eta_1(x_k'),\ \frac{k}{2}\right) \tag{6·143}$$

设随机变量 W 服从二项分布 $B(k,\ \eta_1(x))$，定义

$$t_k(x)\triangleq\eta_1(x)P\left(W<\frac{k}{2}\right)+\eta_2(x)P\left(W>\frac{k}{2}\right)$$

$$+\frac{1}{2}P\left(W=\frac{k}{2}\right)$$

$$=\eta_1(x)\sum_{0<i<k/2} b(k;\eta_1(x),\ i)$$

$$+\eta_2(x)\sum_{\frac{k}{2}<i<k} b(k;\eta_1(x),\ i)$$

$$+\frac{1}{2}b\left(k;\ \eta_1(x),\ \frac{k}{2}\right). \tag{6·144}$$

使用引理6·5 即有

$$|r_n(x, x_1', \cdots, x_k') - t_k(x)| \leqslant \frac{3}{2}\sum_{i=1}^{k} |\eta_1(x_i') - \eta_1(x)| \tag{6·145}$$

若假定 $\eta_1(x)$ 连续，则 $\eta_2(x) = 1 - \eta_1(x)$ 亦然.仿定理6·18中（6·135）式的证法可知：对几乎所有的 x 及每一 $1 \leqslant i \leqslant k$，有

$$X_i'(x) \longrightarrow x, \text{ a.s. 当 } n \to \infty,$$

其中 $X_i'(x)$ 为当 $X = x$ 时 X_i' 的标记。从而由（6·145）可得：对几乎所有的 x

$$r_n(x; X_1', \cdots, X_k') \longrightarrow t_k(x), \text{a.s. 当 } n \to \infty. \tag{6·146}$$

再使用控制收敛定理即得下述的定理。

定理6·20 设 $M = 2$，$\eta_1(x)$ 连续，则

$$\lim_{n\to\infty} R_n^{(k)} \triangleq \lim_{n\to\infty} E(r_n(X; X_1', \cdots, X_k')) = Et_k(X). \tag{6·147}$$

为明瞭（6·147）式的意义，让我们进一步考察 $t_k(x)$。

记 $\eta(x) = \min(\eta_1(x), \eta_2(x))$,

$$b(k, p, i) = \binom{k}{i} p^i (1-p)^{k-i},$$

$$B(k, p, i) = \sum_{j=0}^{i} b(k, p, j),$$

$$B_k = B\left(k; \eta(x), \frac{k}{2}\right) + \frac{1}{2} b\left(k; \eta(x), \frac{k}{2}\right)$$

$$= \sum_{0 \leqslant i < k/2} b(k; \eta(x), i) + \frac{1}{2} b\left(k; \eta(x), \frac{k}{2}\right), \tag{6·148}$$

则

$$t_k(x) = \eta(x) B_k + (1 - \eta(x))(1 - B_k)$$
$$= 1 - \eta(x) - B_k(1 - 2\eta(x)). \tag{6·149}$$

我们可以证明

$$t_1(x) = t_2(x) \geqslant t_3(x) = t_4(x) \geqslant \cdots \tag{6·150}$$

而且

$$\lim_{k\to\infty} t_k(x) = r^*(x). \qquad (6\cdot151)$$

事实上，由 B_k 的定义不难断定

$$B_1 = B_2 \leqslant B_3 = B_4 \leqslant \cdots, \qquad (6\cdot152)$$

$$\lim_{k\to\infty} B_k = 1 \text{ 当 } \eta(x) < \frac{1}{2}; = 1/2 \text{ 当 } \eta(x) = 1/2. \quad (6\cdot153)$$

因 $\eta(x) \leqslant 1/2$，由（6·149）知 $t_k(x) \to \eta(x)$。再由 $r^*(x) = \eta(x)$ 即得（6·151）。(6·150)式则易由（6·149）与（6·152）推出。

再用控制收敛定理得到

$$\lim_{k\to\infty} Et_k(X) = Er^*(X) = R^*. \qquad (6\cdot154)$$

由此可对定理 6·20 的结论作如下解释：k- N.N. 规则的错判概率当样本容量无限增大时有一个不低于 R^* 的同 k 有关的极限。而此极限随着 k 增大而任意接近 Bayes 风险 R^*。事实上，当 k 无限增大时已无须依对 x 的距离重新排序，判别实际上直接基于 $X_1, \cdots, X_n$ 作出；再随着 n 无限增大其功效当然应与 Bayes 判别相同。

另一个有实际意义的问题是：虽然 $Et_k(X) \geqslant R_*$，但对固定的 k，对所有(X, Y)的可能分布，比值 $Et_k(X) / R^*$ 的上界是多少？有了这个上界，在某种意义上可以定量地考察 k-N.N. 法的效用，而且显然这个界同 k 有关，因而这个界对 k 的选择也有所帮助。记

$$T_k = \sup_{(X,\ Y)\text{的分布}} (Et_k(X))/R^*,$$

已知有下述结果：

当 $k=1$，2 时，$T_k \leqslant 1 + \sqrt{\dfrac{2}{k}}$，

当 $k \geqslant 3$ 时，$T_k \leqslant 1 + \sqrt{\dfrac{1}{k}}$，

当 $k \geqslant 5$ 时且为奇数时，$T_k \leqslant 1 + \dfrac{\alpha\sqrt{k}}{k-3.25}\left(1 + \dfrac{\beta}{\sqrt{k-3}}\right)$，

其中

$$\alpha\approx 0.3399,\ \beta\approx 0.9749.$$

五、N.N. 法在预测中的应用

在§6·3的四中已提出非参数预测问题，我们仍沿用那里所使用的记号。设因变量Y为一维的，自变量X为d维的，$Z_n=\{(X_1, Y_1), \cdots, (X_n, Y_n)\}$为来自$(X, Y)$的独立同分布样本，$L$为给定的损失函数，用$\delta(x)$表示预测规则，$\delta^*(x)$为Bayes预测，对任给预测$\delta$，若记$F$为$X$的边缘分布函数，$F(\cdot|x)$为给定$X=x$时，$Y$的条件分布函数，则有

$$R(\delta)=E[E(L(Y, \delta(X))|X)]$$
$$=\int\left[\int L(y, \delta(x))F(dy|x)\right]dF(x),$$

记

$$r(x)\triangleq r(\delta;x)=E[L(Y, \delta(X))|X=x]$$
$$=\int L(y, \delta(x))F(dy|x),$$

称$r(\delta;x)=r(x)$为给定$X=x$时δ的后验预测风险，相应的Bayes预测δ^*的后验风险记为$r^*(x)$。易知Bayes预测是使后验风险达到最小的预测规则。现(X, Y)的分布未知，因而无法使用后验风险达到极小的办法求得一个预测。本段采用N.N. 法构造基于Z_n的预测，务求其风险（或后验风险）尽可能地接近R^*(或$r^*(x)$)。仍用X_n'记$X_1,\cdots,X_n$中与X最接近者，而与之匹配的记为Y_n'。当$X=x$时，记相应的Y_n'为$Y_n'(x)$。定义最近邻预测为：

当$X=x$时，用$Y_n'(x)$预测Y.

记

$$r_n(x, X_n')=E[L(Y, Y_n')|X=x, Z_n]$$
$$=E[L(Y, Y_n')|X=x, X_n'], \quad (6\cdot155)$$

$$r_n(x)=E[L(Y, Y_n')|X=x], \quad (6\cdot156)$$

它们分别是N.N. 预测在给定$X=x$，X_n'及给定$X=x$时的后

验风险．再记

$$R_n \triangleq EL(Y,\ Y'_n), \tag{6·157}$$

R_n 是 N.N. 预测的（无条件）风险，用以刻划 N.N. 预测的好坏程度。而 $r_n(x)$ 则表示在给定预测点 $X=x$，反复使用 N.N. 预测其条件平均损失。至于 $r_n(x,\ X'_n)$ 可作相同的解释。因此在实际使用时，$r_n(x)$ 及 $r_n(x,\ X'_n)$ 更适合操作人员的要求。但 R_n 在理论分析时要用到．显然上述三个量有以下关系

$$R_n = Er_n(X),\quad r_n(x) = Er_n(x,\ X'_n). \tag{6·158}$$

由于平方损失 $L(y,\ a)=(y-a)^2$ 是常用且方便的一种损失函数，我们限定在平方损失下讨论 N. N. 预测的性质．已知在平方损失下其 Bayes 预测为给定 X 时 Y 的条件期望。若假定 $E(Y)^2<\infty$，则 δ^* 的后验风险为给定 X 时 Y 的条件方差，即

$$\begin{aligned} r^*(x) &= \mathrm{Var}(Y|X=x) \\ &= E(Y^2|X=x) - [E(Y|X=x)]^2, \end{aligned} \tag{6·159}$$

记

$$\mu_1(x)=E(Y|X=x),\quad \mu_2(x)=E(Y^2|X=x). \tag{6·160}$$

我们有下述的

定理6·21 若 $\mu_i(x)$ 连续，$i=1,\ 2$，则对几乎每一 x，有

$$\lim_{n\to\infty} r_n(x,\ X'_n) = 2r^*(x),\ \text{a.s.} \tag{6·161}$$

证明 易知

$$r_n(x,\ x'_n) = \mu_2(x) - 2\mu_1(x)\mu_2(x'_n) + \mu_2(x'_n)$$

因 $X'_n \longrightarrow x$，a.s.对几乎每一 x 成立，以及 μ_1、μ_2 连续，即有对几乎每一 x

$$\lim_{n\to\infty} r_n(x,\ X'_n) = 2[\mu_2(x) - \mu_1^2(x)],\ \text{a.s.}$$

再由 $r^*(x)$的表达式（6·159）即得证（6·161），定理证毕。

注意到上面的定理暗含着假定 $E(Y)^2<\infty$，不然的话 $r^*(x)$ 无意义．此时可推出 $r_n(x)=Er_n(x,\ X'_n)$ 存在．但由（6·161）还不能得出 $\lim\limits_{n\to\infty} r_n(x)$ 的存在性。

定理6·22 设 $E(Y)^2<\infty$，$\mu_1(x)$ 连续（在 ρ 下）。若距离 ρ 使得 $E\rho^2(X, 0)<\infty$，且存在绝对常数 A、B 有

$$|\mu_1(x)-\mu_2(y)|\leqslant A\rho(x, y)$$

$$|\sigma^2(x)-\sigma^2(y)|\leqslant B\rho^2(x, y)$$

其中 $\sigma^2(x)=\mathrm{Var}(Y|X=x)$，

则有

$$\lim_{n\to\infty} r_n(X)=2r^*(X), \text{ a.s.}, \tag{6·162}$$

而且

$$\lim_{n\to\infty} R_n=2R^*. \tag{6·163}$$

证明 由假设可知

$$\begin{aligned} r_n(x, X'_n) &= \sigma^2(x)+\sigma^2(X'_n)+(\mu_1(x)-\mu_1(X'_n))^2 \\ &= 2\sigma^2(x)+(\sigma^2(X'_n)-\sigma^2(x)) \\ &\quad +(\mu_1(x)-\mu_1(X'_n))^2 \\ &\leqslant 2\sigma^2(x)+B\rho^2(x, X'_n)+A^2\rho^2(x, X'_n) \\ &= 2\sigma^2(x)+(A^2+B)\rho^2(x, X'_n) \end{aligned}$$

但 $\rho^2(x, X'_n)\leqslant\rho^2(x, X_1)\leqslant 2\rho^2(x, 0)+2\rho^2(X_1, 0)$，因而由（6·161）式，依控制收敛定理即有

$$\lim_{n\to\infty} r_n(x)=\lim_{n\to\infty} Er_n(x, X'_n)=2r^*(x), \text{对 a.e. } x, \tag{6·164}$$

此即（6.162）式成立。又对每一 x，有

$$r_n(x)=Er_n(x, X'_n)\leqslant 2\sigma^2(x)+2(A^2+B)[\rho^2(x, 0)+E\rho^2(X_1, 0)],$$

及 $E\sigma^2(X)\leqslant E(Y)^2+E(E(Y|X))^2\leqslant 2E(Y)^2<\infty$，

$E\rho^2(X, 0)<\infty$，

由（6·164）再次使用控制收敛定理得到（6·163）式，定理证毕。

关于（6·163）的解释，定理6·18后面的注解同样可适用于此，只须将那里的“判别”换成这里的“预测”就行了。

下面的例子说明极限 $\lim\limits_{n\to\infty} R_n$ 可以不存在。

例6·6 设X有 Cauchy 分布，给定$X=x$，Y 的条件分布为正态 $N(x, 1)$，此时

$$\mu_1(x)=x,\ \mu_2(x)=1+x^2,\ \sigma^2(x)=1,$$

因 $E|X|=\infty$，故一般说来定理 6·22 中的条件 $E\rho^2(X, 0)<\infty$ 不成立。易知

$$r_n(x, X_n')=2+(x-X_n')^2, \tag{6·165}$$

因而

$$R_n=Er_n(X, X_n')=2+E(X-X_n')^2. \tag{6·166}$$

对任给 $a>0$，当 $x>a$ 时，

$$P(|x-X_n'|\geqslant a|X=x)\geqslant P(X_1<0, \cdots, X_n<0)=\left(\frac{1}{2}\right)^n.$$

于是

$$P(|X-X_n'|\geqslant a)=EP(|X-X_n'|\geqslant a|X)$$
$$\geqslant\int_{\{x\geqslant a\}}P(|x-X_n'|\geqslant a|X=x)dF(x)$$
$$\geqslant\left(\frac{1}{2}\right)^n P(X\geqslant a),$$

其中 F 为 Cauchy 分布的分布函数。由上述不等式经简单计算可知：存在一仅同n有关（与a无关）的常数 $c>0$，使得当a充分大时有

$$P(|X-X_n'|\geqslant a)\geqslant c/a,$$

因而当a充分大时，有

$$E(X-X_n')^2\geqslant a^2P(|X-X_n'|\geqslant a)\geqslant ac.$$

于是 $R_n=2+E(X-X_n')^2=+\infty$，自然无从谈起$\{R_n\}$的极限。但有趣的是，可以证明：

$$r_n(x)=2+E(x-X_n')^2<\infty, \text{ 对所有 } x,$$

而且对每一x，有

$$\lim_{n\to\infty} r_n(x)=2r^*(x).$$

习　　题

6-1　设 $f(x)$ 是 R 上的概率密度，若 f 在 R 上一致连续则 f 在 R 上有界，且 $\lim_{|x|\to\infty} f(x)=0$.

6-2　证明引理6·2.

6-3　证明（6·17）式。

6-4　设 $K(u)$ 是正态 $N(0,1)$ 密度，f 为正态 $N(\mu,\sigma^2)$ 密度。$X_1,\cdots,X_n$ 是来自 f 的 iid. 样本. f_n 是基于 $X_1,\cdots,X_n$ 的具核 K 及窗宽 h_n 的核估计。求 $Ef_n(x)$ 及 $\mathrm{Var}(f_n(x))$。

6-5　随机数的模拟　设 $X_1,\cdots,X_n$ iid.，X_1 有未知密度 f. f_n 是基于 $X_1,\cdots,X_n$ 具核 K 及窗宽 h_n 的核估计。今从1，2，…，n 中随机抽取一个记为 I；ε 是与 $X_1,\cdots,X_n$ 独立具密度 K 的随机变量。则对给定 $X_1,\cdots,X_n$，$Y=X_I+h_n\varepsilon$ 的条件密度为 f_n.

6-6　设 $X_1,\cdots,X_n$ 是来自未知密度 f 的 iid. 样本，$h_n>0$ 是给定常数序列，满足 $\lim_{n\to\infty} h_n=0$. 记

$$N_n(a,b)={}^{\#}(\{i:X_i\in(a,b),\ i=1,2,\cdots,n\}),$$

定义

$$f_n(x)=N_n(x-h_n,\ x+h_n)/2nh_n,$$

则有

(1) $\lim_{n\to\infty} Ef_n(x)=f(x)$，$x\in C(f)$，

(2) 若 $\lim_{n\to\infty} nh_n=\infty$，则当 $n\to\infty$ 时

$$f_n(x)\xrightarrow{P}f(x),\quad x\in C(f).$$

6-7　（续6题）　设 f 在 R 上一致连续，若

$\lim_{n\to\infty} h_n=0$，$\lim_{n\to\infty} nh_n^2/(\log n)=\infty$，

则

$$\lim_{n\to\infty}\sup_x|f_n(x)-f(x)|=0,\ \text{a.s.}$$

6-8 设 $X_1, \cdots, X_n$ iid., X_1 有分布 F（并不假定 F 有密度），f_n 是基于 $X_1, \cdots, X_n$ 具核K及窗宽 h_n 的核估计。若假定K在R上有界变差，

$$\lim_{n\to\infty} h_n=0, \quad \lim_{n\to\infty} nh_n^2/(\log n)=\infty,$$

则

$$\lim_{n\to\infty} \sup_x |f_n(x)-Ef_n(x)|=0, \text{ a.s.}.$$

6-9 在上题的假设条件下，若

$$\lim_{n\to\infty} \sup_x |f_n(x)-g(x)|=0, \text{ a.s. 对某个 } g,$$

则 F是处处连续的。

6-10 设 f_n 是未知密度f的核估计，$\lim\limits_{n\to\infty} h_n=0$，则

$$\lim_{n\to\infty} E_f\left(\int|f_n(x)-f(x)|dx\right)=0, \text{对任何 } f,$$

又若 $\int K^2(y)dy<\infty$，则当 $\lim\limits_{n\to\infty} nh_n=\infty$时，有

$$\lim_{n\to\infty} \int \operatorname{Var}_f(f_n(x))dx=0, \text{对任何 } f.$$

6-11 设 F 为一维分布函数，$p\in(0,1)$．令

$$c=\sup\{t:F(t)\leqslant p\}, \quad d=\inf\{t:F(t)\geqslant p\}$$

则

(1) $-\infty<d\leqslant c<+\infty$;

(2) ξ_p 为 F 的 p 分位数当且仅当 $\xi_p\in[d, c]$.

6-12 求证在平方损失下，Bayes 预测为 $E(Y|x)$；在绝对值损失下为 $\xi\left(\frac{1}{2}|x\right)$（即条件中位数）。

6-13 设K是有界的具紧支撑的概率密度．(X, Y), (X_1, Y_1), $\cdots$, (X_n, Y_n) iid., $\eta_i(x)=P(Y=i|X=x)$, $i=1, \cdots, M$ 未知．构造 $\eta_i(x)$ 的核估计为

$$\eta_{ni}(x)=\sum_{j:Y_j=i} K\left(\frac{x-X_j}{h_n}\right)\Big/\sum_{j=1}^n K\left(\frac{x-X_j}{h_n}\right), i=1,\cdots,M.$$

若 $\lim\limits_{n\to\infty} h_n=0$，$\lim\limits_{n\to\infty} nh_n^d=\infty$，则对任何 $r\geqslant 1$

$$\lim_{n\to\infty}\sum_{i=1}^{M} E|\eta_{ni}(X)-\eta_i(X)|^r=0.$$

6-14 设 Y_1，Y_2，…，X，X_1，X_2，…，相互独立，$\{W_{ni}\}$ 是由 X_1，…，X_n 所确定的权函数，且诸 Y_i 服从 $N(0, 1)$ 分布．若 $\sum\limits_{i=1}^{n} W_{ni}(X)Y_i \xrightarrow{P} 0$，当 $n\to\infty$，则 $\sum\limits_{i=1}^{n} W_{ni}^2(X) \xrightarrow{P} 0$，当 $n\to\infty$．

6-15 设 $M=2$，$p_1=p_2=1/2$，$f_1(x)=\begin{cases}2(1-x), & 0<x<1\\ 0, & \text{其它},\end{cases}$ $f_2(x)=\begin{cases}2x, & 0<x<1\\ 0, & \text{其它}.\end{cases}$
R_n 是 N.N. 判别的错判概率，则 $\lim\limits_{n\to\infty} R_n=1/3$.

6-16 设 k 为正整数，$0<p\leqslant 1/2$，$t_k=1-p-B_k(1-2p)$，$B_k=\sum\limits_{0\leqslant i<k/2} b(k;\ p,\ i)+\frac{1}{2}b\left(k;\ p,\ \frac{k}{2}\right)$．则有

(1) $t_1=t_2\geqslant t_3=t_4\cdots$；(2) $\lim\limits_{k\to\infty} t_k=p$.

6-17 设 (X, Y)，(X_1, Y_1)，…，(X_n, Y_n) iid.，$\eta_i(x)=P(Y=i|X=x)$，$i=1, 2, \cdots, M$，F 为 X 的分布函数．定义 $\eta_i(x)$ 的估计如 $\eta_{ni}(x)$ 如问题13，判别规则为

当 $\eta_{ni}(x)=\max\limits_{1\leqslant j\leqslant M}\eta_{nj}(x)$ 时，判 x 为 i 类．

记此规则为 $g_n(x)$，$L_n=p(g_n(X)\neq Y|X_n)$，R^* 为 Bayes 规则的错判概率($X_n=\{(X_1, Y_1), \cdots, (X_n, Y_n)\}$)．则有

$$0\leqslant L_n-R^*\leqslant 2\sum_{i=1}^{M}\int|\eta_{ni}(x)-\eta_i(x)|dF(x).$$

6-18 设 X_1'，…，X_k' 为 X_1，…，X_n 中与 X 最接近的前 k 个点，相应的 Y 记为 Y_1'，…，Y_k'．用 $\hat{Y}_n=\frac{1}{k}\sum\limits_{i=1}^{k} Y_i'$ 预测 Y，L 为平方损失．假定 $Y\sim N(\mu, \sigma_1^2)$，$X|Y=y\sim N(y, \sigma_2^2)$．求：

(1) Bayes 预测 δ^* 及风险 R^*，

(2) $r_n(x;\ x_1',\cdots,x_k')=p(\hat{Y}_n\neq Y\,|\,X=x,X_i'=x_i',$
$i=1,\cdots,k)$.

6-19 设 $X, X_1, \cdots, X_n$ iid.，给定 $k\geqslant 1$ 及 $X=x$，Z_1'，…，Z_k' 如上题所示。则对几乎所有 x，有

$$\max_{1\leqslant i\leqslant k} \rho(x,\ X_i') \xrightarrow{p} 0,\ 当 n\to\infty.$$

6-20 在例6·6中证明：$\lim_{n\to\infty} r_n(x)=2r^*(x)$。

习 题 提 示

第 二 章

2-2 只须证明：若F密度不存在（非绝对连续），则$X_{(r)}$的密度不可能存在。事实上，若F非绝对连续，则存在Lebesgue零测集A，使$F(A)=P(X\in A)>0$．这时，$P(X_i\in A, i=1, \cdots, n)=F^n(A)>0$．因此$P(X_{(r)}\in A)\geqslant P(X_i\in A, i=1, \cdots, n)>0$．这表明$X_{(r)}$之分布非绝对连续。

2-4 若F在某点a有跳跃$p>0$，则$F(X)$取$F(a)$为值的概率$\geqslant p>0$．从而$F(X)$不可能有均匀分布。

2-5 根据G的定义，去证明对任何$x\in(0, 1)$，有$G(U)\leqslant x\Longleftrightarrow U\leqslant F(x)$（$F$右连续）。

2-6 $V_{(1)}, \cdots, V_{(n+1)}$同分布通过直接计算$U_{(i)}-U_{(i-1)}$的分布（$i=2, \cdots, n$，$V_{(n+1)}$单独算）即可证实。其不独立可以从$V_{(1)}+\cdots+V_{(n+1)}=1$看出。任一对不独立则从其和$\leqslant 1$看出，因每一个都可在$(0, 1)$内取值。

如果把$(0, 1)$折成一半径为$\frac{1}{2\pi}$的圆周，则由对称性考虑，n个点分割成的$n+1$个圆弧位置完全平等，由这就不难证明其同分布性。

2-7 不妨设$\mu=0$．利用$X\stackrel{d}{=}-X$，不难推出$X_{(r)}\stackrel{d}{=}-X_{(n+1-r)}$．由此即推出$E(\hat{m})=0$．直接计算也以设$\mu=0$为方便（$\mu\neq 0$时，以$X_i-\mu$代$X_i$即可）。

2-9(a) 为证$E(R)\to\infty$，不妨设X无上界。这意思是说$P(X>c)>0$对任何实数c．取a，使$F(a)>0$．给定$A>0$．则$p=P(X\geqslant 2A+a)>0$.

有$P(X_{(1)}\leqslant a)=1-(1-F(a))^n\to 1$，当$n\to\infty$。

又 $P(X_{(n)}\geqslant 2A+a)=1-(1-p)^n\to 1$，当 $n\to\infty$. 故

$$P(X_{(1)}\leqslant a,\ X_{(n)}\geqslant 2A+a)\to 1,\ 当\ n\to\infty$$

因此当 n 充分大时，$P(R\geqslant 2A)\geqslant P(X_{(1)}\leqslant a,$ $X_{(n)}\geqslant 2A+a)\geqslant\frac{1}{2}$，而 $E(R)\geqslant 2A\cdot\frac{1}{2}\geqslant A$. 因 A 任意，知 $E(R)\to\infty$.

(b) 为证 $E(R)\to\sup X-\inf X$，分别证 $E(X_{(n)})\to\sup X$，$E(X_{(1)})\to\inf X$. 以前者为例，按 $\sup X$ 的定义，对任给 $\varepsilon>0$，仿 a 之证法，不难证得 $P(X_{(n)}\geqslant\sup X-\varepsilon)\to 1$，这一事实，结合 $X_{(n)}$ 有界及 $X_{(n)}\leqslant\sup X$，即得 $E(X_{(n)})\to\sup X$.

(c) 欲证 $E(R_n)$ 严增（此处以 R_n 记 $X_{(n)}-X_{(1)}$），注意 $X_{(n+1)}=\max(X_{(n)},X_{n+1})\geqslant X_{(n)}$. 由此知 $E(R_n)$ 随 n 非降。欲证其严增，只须证 $P(X_{n+1}>X_{(n)})>0$. 为此，利用 X 非退化，可找到 a，使 $P(X<a)>0$，$P(X>a)>0$. 但 $P(X_{n+1}>X_{(n)})\geqslant P(X_{n+1}>a,\ X_{(n)}<a)=P(X_{n+1}>a,$ $X_1<a,\cdots,X_n<a)=P^n(X<a)P(X>a)>0$. 得证。

2-12　设 X 有两个不同的对称中心。不失普遍性设其一为 0，另一为 $a\neq 0$，故 X 及 $X-a$ 都关于 0 对称，于是有

$X\overset{d}{=}-X$，$X-a\overset{d}{=}-(X-a)=-X+a$. 由后一式知

$X\overset{d}{=}-X+2a$. 此与第一式结合，得 $-X\overset{d}{=}-X+2a$，而 $a\neq 0$. 这不可能（为什么？）

2-13　考察负指数密度 $e^{-x}I(x>0)$.

2-14　前半见第 7 题。相合与否取决于中位数是否唯一。若唯一则相合，否则不相合（考虑样本大小为奇数时的情况）。而且，如果中位数不唯一，若以 m_n 记 $X_1,\cdots,X_n$ 的样本中位数，序列 $\{m_{2n}:n=1,2,\cdots,\}$ 也不是对称中心的相合估计。请证明之。

2-16　此是下面容易证明的结果的推论：设有三串随机变量 $\{\xi_{in}:n=1,2,\cdots\}$，$i=1,2,3$，$\xi_{1n}\leqslant\xi_{2n}\leqslant\xi_{3n}$ 对一切 n. 若

ξ_{1n} 和 ξ_{3n} 当 $n\to\infty$ 时都依分布收敛于同一分布 F，则 ξ_{2n} 也依分布收敛于 F. 先证明这一事实。从表达式（2·16）定出以上三串随机变量，再利用定理 2·2.

2-21 T_1 非完全可证明如下。定义 $g(T)=I(X_1\leqslant 1)-I(X_2\leqslant 1)$. 则 $E_F(g(T))=0$，对任何 $F\in\mathscr{F}$. 但 $g(T)$ 并不以概率 1 为 0. 其余容易。

2-22 充分性证明与定理 2·6 同。完全性可通过取 $I(Y_1\leqslant 1)-I(Z_1\leqslant 1)$，按上题方法证明之。

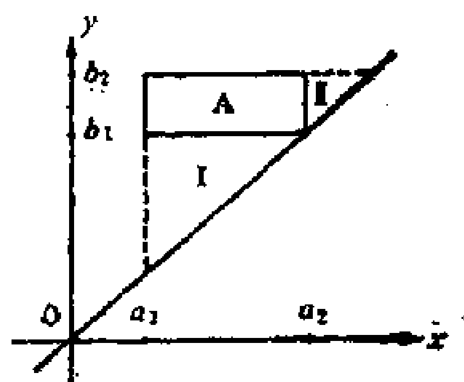
2-23 记 $T=(X_{(1)},\cdots,X_{(n)})$. 当 $n\geqslant 3$ 时，取

$g_1(T)=\dfrac{X_{(2)}-X_{(1)}}{X_{(n)}-X_{(1)}}$，证明 $Eg_1(T)=a$ 与 θ_1，θ_2 无关，

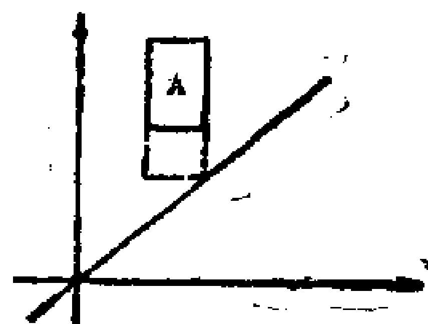
然后取 $g(T)=g_1(T)-a$. 当 $n=2$ 时，为证 $(X_{(1)},X_{(2)})$ 完全，只须证明：若 $g(x,y)$ 定义于 $\{x<y\}$ 上，且

$$\iint\limits_{\theta_1<x<y<\theta_2} g(x,y)\,dxdy=0,\ 对一切\ \theta_1<\theta_2,$$

则 $g(x,y)$ 在集合 $\{x<y\}$ 上为 0. 为证后者，只须证明：若 $A=\{(x,y):a_1<x<a_2,\ b_1<y<b_2\}$，此处 $-\infty<a_1<a_2<\infty$，$-\infty<b_1<b_2<\infty$ 且 $A\subset\{x<y\}$，则有 $\iint\limits_A g(x,y)\,dx\,dy=0$. 为证此，先考虑图一的情况，这相当于 $a_2=b_1$ 的情况，这时矩形 A 等于图中那个大三角形减去两个小三角形 I 和 II. 因由假定，g 在这三个三角形上的积分皆为 0，故 $\iint\limits_A g(x,y)\,dxdy=0$，一般情况如图二，其中 A 等于两个矩形之差，这



图一



图二

两个矩形都属于前一情况。

第 三 章

3-1 (a)如果最小方差无偏估计$\hat{\theta}$存在，则因$\bar{X}$也是无偏估计，对一切对称分布F应有$\mathrm{Var}_F(\bar{X})\geqslant\mathrm{Var}_F(\hat{\theta})$. 特别，因正态分布为对称分布，上式对一切正态分在成立，由此将推出：$\hat{\theta}$是正态分布期望μ的最小方差无偏估计。但在估计理论中已证明：μ的最小方差无偏估计为$\bar{X}$且唯一，这证明$\hat{\theta}$必须是$\bar{X}$。但$\bar{X}$并非最小方差无偏估计，因为若以m_n记样本中位数，则m_n也是无偏估计，而对某些对称分布F（试举一例），m_n的方差小于$\bar{X}$的方差。

(b) 问题出在对所说的分布族而言，次序统计量$(X_{(1)},\cdots,X_{(n)})=T$并非完全的。此可由当分布对称时，有$X_{(n)}-X_{(n-1)}\overset{d}{=}X_{(2)}-X_{(1)}$，用第二章22题的方法证明之。

3-2 如方差的级为1，则存在$g(x_1)$，使$\mathrm{Var}(F)=E_F(g(X_1))$. 特别，对$F$为$(0,\theta)$区间均匀分布成立，$\theta>0$。于是有$\frac{1}{\theta}\int_0^\theta g(x)\,dx=\frac{1}{12}\theta^2$，对一切$\theta>0$. 由此知

$$\int_0^\theta g(x)\,dx=\frac{1}{12}\theta^3,\ g(\theta)=\frac{1}{4}\theta^2,\ \theta>0.$$

但对这样的g，若分布F取9和10的概率都是$\frac{1}{2}$，将不满足

$\mathrm{var}(F)=E_F(g(X_1))$. 于是知这种$g$不存在。

3-3(a) 设X，Y独立同分布，有公共分布F，则有

$\int_{-\infty}^{\infty}F(x)dF(x)=P(X\leqslant Y)$（按右连续）。因$X\overset{d}{=}Y$，有

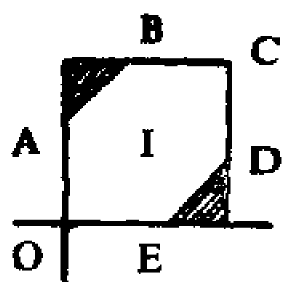
$$\begin{aligned}1&=P(X<Y)+P(X>Y)+P(X=Y)\\&=2P(X<Y)+P(X=Y)\qquad(*)\end{aligned}$$

此因$P(X<Y)=P(X>Y)$. 由第一式，知

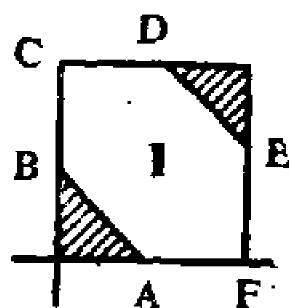
$$P(X \leqslant Y) = P(X<Y) + P(X=Y) \geqslant \frac{1}{2}$$

等号仅当 $P(X=Y)=0$ 成立，而后一事实当且仅当 F 处处连续才对（请证明）。由（＊）的第二式，及 F 左连续时 $\int_{-\infty}^{\infty} F(x)\,dF(x) = P(X<Y)$，知 $\int_{-\infty}^{\infty} F(x)\,dF(x) \leqslant \frac{1}{2}$，等号当且仅当 $P(X=Y)=0$ 时成立。

(b) $\iint\limits_{-\infty}^{\infty} F(x,y)\,dF(x,y)$ 介于 0 与 $\frac{1}{2}$ 之间可由此积分等于 $P(X_1 \leqslant X_2,\ Y_1 \leqslant Y_2)$，用上题方法去证明。此处 $(X_1,\ Y_1)$，$(X_2,\ Y_2)$ 独立同分布，有公共分布 $F(x,\ y)$。此积分之值可为 $[0,\ \frac{1}{2}]$ 之间任何数的断言，可通过取 F 为一些具体分布去证明。例如，取图一的多边形 I 上的均匀分布为 F（I 为单位正方



图三



图四

形的一部分，且 $OA=BC=CD=OE=a$），令 $OA=a$ 在 $[0,1]$ 内变化（当 $a=0$ 时，F 成为对角线 OC 上的均匀分布），可得上述积分在 $[\frac{1}{4},\ \frac{1}{2}]$ 内任何值。用图二中的多边形 II 上的均匀分布为 F（结构与图一相似），可得上述积分在 $[0,\ \frac{1}{4}]$ 内任何值。

(c) 记 $G(x) = \lim\limits_{y \to x-} F(y)$，即 F 在 x 点的左极限。因 $P(X<Y) = E\{P(X<Y|Y)\} = \int_{-\infty}^{\infty} G(x)\,dF(x-\theta) = \int_{-\infty}^{\infty} G(x+\theta)\,dF(x)$．故

$$P(X<Y)-\int_{-\infty}^{\infty}F(x)dF(x)$$
$$=\int_{-\infty}^{\infty}[G(x+\theta)-F(x)]dF(x),$$

分两种情况：1. F 有一个跳跃点 a，其跃度 $p>0$. 则
$$\int_{-\infty}^{\infty}[G(x+\theta)-F(x)]dF(x)$$
$$\geqslant[G(a+\theta)-F(a)]F(\{a\})\geqslant p\cdot p=p^2>0$$
故 $P(X<Y)\geqslant p^2+\int_{-\infty}^{\infty}F(x)dF(x)\geqslant p^2+\frac{1}{2}>\frac{1}{2}$. 2. F处处连续。这时 $G(x)=F(x)$. 存在点 a，使当 $x>a$ 时总有 $F(x)>F(a)$. 这时取 $\varepsilon>0$ 充分小，有
$$\int_{-\infty}^{\infty}[G(x+\theta)-F(x)]dF(x)$$
$$\geqslant\int_{a}^{a+\varepsilon}[F(x+\theta)-F(x)]dF(x),$$
因 $F(a+\theta)-F(a)>0$ 而 F 连续，故当 $\varepsilon>0$ 充分小时，有 $b\equiv\inf\{F(x+\theta)-F(x):a\leqslant x\leqslant a+\varepsilon\}>0$. 又 $F(a+\varepsilon)>F(a)$，故
$$\int_{a}^{a+\varepsilon}[F(x+\theta)-F(x)]dF(x)\geqslant b[F(a+\varepsilon)-F(a)]>0,$$
其余与情况 1 一样。

3-5 将题中 $\hat{\theta}(F)$积分号下的平方展开得三项。第一项为
$$\int_0^{\infty}\int_0^{\infty}\bar{F}^2(s)\bar{F}^2(t)dF(s)dF(t)$$
$$=\int_0^1\int_0^1(1-x)^2(1-y)^2dxdy=\frac{1}{9},$$不须估计.第二项为
$$-2J=-2\int_0^{\infty}\int_0^{\infty}\bar{F}(s)\bar{F}(t)\bar{F}(s+t)dF(s)dF(t)$$
证明：若 $X_1,\ \cdots,\ X_6\sim F$，则
$$J=P(X_1>X_4+X_5,\ X_2>X_4,\ X_3>X_5)$$
第三项为$\int_0^{\infty}\int_0^{\infty}\bar{F}^2(s+t)dF(s)dF(t)=P(X_1>X_3+X_4,$

$X_2>X_3+X_4$）．利用这些事实，不难找到一个核作为 $\theta(F)$ 的无偏估计（依赖 $X_1, \cdots, X_8$），由此通过作 U 统计量即可。

3-6 前一问简单．后一问可通过取具体分布去算。例如第三章习题提示 3(b)中之分布。

3-7 问题在于找出 $\theta(F)$ 的一个无偏估计。以 (X_1, Y_1)，(X_2, Y_2)，…记从 F 中抽出的简单样本，先证明

$$\iint_{-\infty}^{\infty} F^2(x, y)dF(x, y)$$

$$=P(X_1 \leqslant X_3,\ X_2 \leqslant X_3,\ Y_1 \leqslant Y_3,\ Y_2 \leqslant Y_3),$$

$$\iint_{-\infty}^{\infty} F(x, y)F_1(x)F_2(y)dF(x, y)$$

$$=P(X_1 \leqslant X_4,\ Y_1 \leqslant Y_4,\ X_2 \leqslant X_4,\ Y_3 \leqslant Y_4),$$

$$\iint_{-\infty}^{\infty} F_1^2(x)F_2^2(y)dF(x, y)$$

$$=P(X_1 \leqslant X_5,\ X_2 \leqslant X_5,\ Y_3 \leqslant Y_5,\ Y_4 \leqslant Y_5),$$

以此为基础即可构造出 $\theta(F)$ 之无偏估计，再作成 U 统计量即可。

第 四 章

4-1 以 F 记 X 的分布．若 $i=1$，则当 $X_1=x$ 时，为使 X_1 的秩 $R_1=k$，必须在 $X_2, \cdots, X_n$ 中，有 $k-1$ 个小于 x，另 $n-1-k$ 个大于 x．这是一个二项概率 $b(n-1, F(x); k-1)$．于是

$$E(R_1|X_1=x)=\sum_{k=1}^{n}kb(n-1, F(x); k-1)$$
$$=\sum_{i=0}^{n-1} ib(n-1, F(x); i)+\sum_{i=0}^{n-1} b(n-1, F(x), i)$$
$$=(n-1)F(x)+1.$$

若 $i\neq 1$，则情况较复杂。添加条件 $X_i=y$．当 $y<x$ 时有

$$P(R_i=k|X_1=x, X_i=y)=b(n-2, F(y); k-1)$$

若$y>x$，则有

$$P(R_i=k|X_1=x,\ X_i=y)=b(n-2,\ F(y);\ k-2)$$

于是

$$P(R_i=k|X_1=x)=\int_{-\infty}^{x}b(n-2,\ F(y);\ k-1)dF(y)$$

$$+\int_{x}^{\infty}b(n-2,\ F(y);\ k-2)dF(y)$$

由此可得

$$E(R_i|X_1=x)=\int_{-\infty}^{x}(1+(n-2)F(y))dF(y)$$

$$+\int_{x}^{\infty}(2+(n-2)F(y))dF(y)$$

$$=2-F(x)+\frac{1}{2}(n-2)=\frac{n}{2}+1-F(n).$$

4-2 先算$P(R_1=k|X_1=x)$．欲在$X_1=x_1$的条件下有$R_1=k$，必须X样本中恰有i个$<x$，Y样本中恰有j个$<x$，$i+j=k-1$．由此易得

$$P(R_1=k|X_1=x)$$

$$=\sum_{i=0}^{m-1}\binom{m-1}{i}F^i(x)(1-F(x))^{m-1-i}\binom{n}{k-1-i}G^{k-1-i}$$

$$(x)(1-G(x))^{n-k+1+i}$$

再乘以$dF(x)$对x从$-\infty$到∞积分，即得$P(R_1=k)$

4-4(a) L_n的取值只有n个不同的值

$$d_j=-\sum_{i=1}^{n}\frac{i}{n+1}+n\frac{j}{n+1},\ j=1,\ \cdots,\ n$$

各有概率$\frac{1}{n}$（事实上，d_j就是当$R_1=j$时，L_n的取值，而$P(R_1=j)=\frac{1}{n}$）．由此易算出，$(L_n-l_n)/\sigma_n$收敛于均匀分布$R(-\sqrt{3},\ \sqrt{3})$．此例渐近正态性失效的原因是$(C_{n1},\cdots,C_{nn})$不满足条件N．

(b) 此例L_n取值在0与$\sum_{i=1}^{n}e^{(n+1)2/i^2}<2e^{(n+1)^2}$之间，而

$\sigma_n^2 \geqslant \frac{1}{n} \cdot \frac{(n/2)(n/2)}{n}\left(e^{2(n+1)^2} - n\left(\frac{1}{n}2e^{(n+1)^2}\right)^2\right) \geqslant \frac{1}{4}\frac{1}{4}e^{2(n+1)^2}$,

故$\left|\frac{L_n - l_n}{\sigma_n}\right| \leqslant 8$. 因此$(L_n - l_n)/\sigma_n$不可能依分布收敛于$N(0,1)$.

此例渐近正态性失效是因为φ在（0，1）不是平方可积的。

4-5 计算由$a(i) = \left(\frac{i}{n+1}\right)^2$及$\tilde{a}(i) = E(U_{(i)}^2)$（$U_{(1)} \leqslant \cdots \leqslant U_{(n)}$是$R(0,1)$的次序样本）这两个计分函数所产生的标准化线性秩统计量之差，证明它依概率收敛于0即可。

4-6 此概率$P(A)$等于以下四个互斥事件的概率之和：

$A_1 = \{X_1, \cdots, X_n$中有一个为0，其他皆不为0，1$\}$.

$A_2 = \{X_1, \cdots, X_n$中有一个为1，其他皆不为0，1$\}$.

$A_3 = \{X_1, \cdots, X_n$中有一个为0，一个为1，其他皆不为0,1$\}$.

$A_4 = \{X_1, \cdots, X_n$皆不为0，1$\}$.

4-7 $P(\xi = i) = \binom{n}{i}\left(\frac{1}{3}\right)^i\left(\frac{2}{3}\right)^{n-i}, i \geqslant 2$. $P(\xi = 0) = \left(\frac{2}{3}\right)^n + n\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)^{n-1}$.由此算得$E(\xi) = \frac{n}{3}\left(1 - \left(\frac{2}{3}\right)^{n-1}\right)$.

4-8 前一问简单.后一问的证法与第5题相似.此题$(C_{n1}, \cdots, C_{nn})$本为$(0, \cdots, 0, 1, \cdots, 1)$ ($n_1 \uparrow 0$, $n_2 \uparrow 1$).因作线性变换不影响标准化线性秩统计量之值，不妨把$(C_{n1}, \cdots, C_{nn})$取为$\left(-\frac{1}{n_1}, \cdots, -\frac{1}{n_1}, \frac{1}{n_2}, \cdots, \frac{1}{n_2}\right)\left(n_1 \uparrow -\frac{1}{n_1}, n_2 \uparrow \frac{1}{n_1}\right)$.

注意这时$\bar{C}_n$为0. Mood统计量和修改后的统计量分别相应于

$$a_n(i) = \left(\frac{i}{n+1} - \frac{1}{2}\right)^2 \text{(Mood 统计量)}$$

$$\tilde{a}_n(i) = \left(\frac{i}{n} - \frac{1}{2}\right)^2 \text{（修改后统计量）}$$

由它们作成的线性秩统计量分别记为L_n和$\tilde{L}_n$. 有$E(L_n) = E(\tilde{L}_n) = 0$，$\sigma_n^2 = \text{Var}(L_n) = \frac{1}{n-1}\frac{n}{n_1 n_2}\sum_{i=1}^{n}(a_n(i) - \bar{a}_n)^2$,

$\tilde{\sigma}_n^2=\frac{1}{n-1}\frac{n}{n_1n_2}\sum_{i=1}^{n}(\tilde{a}_n(i)-\bar{\tilde{a}}_n)^2$. 因为由定积分定义有

$$\lim_{n\to\infty}\sum_{i=1}^{n}(a_n(i)-\bar{a}_n)^2/n=\lim_{n\to\infty}\sum_{i=1}^{n}(\tilde{a}_n(i)-\bar{\tilde{a}}_n)^2/n$$

$$=\int_0^1\left[\left(x-\frac{1}{2}\right)^2-\frac{1}{12}\right]^2dx=\frac{1}{180}$$

知 $\sigma_n^2/\tilde{\sigma}_n^2\to1$，且 $\sigma_n^2\geqslant\frac{1}{200}\frac{n}{n_1n_2}$，当 n 充分大.故只须证

$\tilde{L}_n/\sigma_n\xrightarrow{\mathscr{L}}N(0,1)$ 即可.注意到

$$\frac{i}{n}-\frac{1}{2}=\frac{n+1}{n}\left(\frac{i}{n+1}-\frac{1}{2}\right)+\frac{1}{2n},$$

易得

$$|L_n-\tilde{L}_n|$$

$$\leqslant\frac{3}{n}\sum_{i=1}^{n_2}\left(\frac{R_i}{n+1}-\frac{1}{2}\right)^2+\frac{n+1}{n^2}\left|\sum_{i=1}^{n_2}\left(\frac{R_i}{n+1}-\frac{1}{2}\right)\right|+\frac{1}{4n}.$$

由于 $L_n/\sigma_n\xrightarrow{\mathscr{L}}N(0,1)$（定理4·4），知

$$\frac{3}{n}\sum_{i=1}^{n_2}\left(\frac{R_i}{n+1}-\frac{1}{2}\right)^2\Big/\sigma_n=\frac{3}{n}L_n/\sigma_n\xrightarrow{P}0,\ \text{当 } n\to\infty, \tag{1}$$

又$\sum_{i=1}^{n_2}\left(\frac{R_i}{n+1}-\frac{1}{2}\right)$为两样本 Wilcoxon 秩和统计量. 因此处有

$n_1\to\infty$，$n_2\to\infty$，有$\sqrt{\frac{12n_1n_2}{n}}\sum_{i=1}^{n_2}\left(\frac{R_i}{n+1}-\frac{1}{2}\right)\xrightarrow{\mathscr{L}}N(0,1)$.

由此及 $\sigma_n^2\geqslant\frac{1}{200}\frac{n}{n_1n_2}$，即知

$$\frac{n+1}{n^2}\left|\sum_{i=1}^{n_2}\left(\frac{R_i}{n+1}-\frac{1}{2}\right)\right|\Big/\sigma_n\xrightarrow{P}0,\ \text{当} n\to\infty, \tag{2}$$

又

$$\frac{1}{4n}\Big/\sigma_n\leqslant\sqrt{200}\sqrt{\frac{n_1n_2}{n}}\frac{1}{n}\leqslant\sqrt{200}/n\longrightarrow0, \tag{3}$$

由(1)、(2)、(3)，知 $|L_n-\tilde{L}_n|/\sigma_n\xrightarrow{P}0$，故 $\tilde{L}_n/\sigma_n$ 与 L_n/σ_n 有同一之极限分布.因后者有极限分布$N(0,1)$，故$\tilde{L}_n/\sigma_n\xrightarrow{\mathscr{L}}N(0,1)$.

4-11 注意事件$\{W=8\}$是以下三个互斥事件之并

$A_1=\{Y_1$，Y_2中有一个秩为5，另一个为3，且X_1,X_2,X_3，中没有与Y_1，Y_2相同的$\}$，

$A_2=\{Y_1$，Y_2中有一个秩为5，剩下一个与X_1，X_2，X_3中的两个相同，还剩下一个X样本比上述样本都小$\}$（例如，$X_2<X_1=X_3=Y_2<Y_1$是一个可能情况），

$A_3=\{Y_1$，Y_2相同且与某一个X样本同，剩下两个X样本都比它们小$\}$（例如，$X_2<X_3<X_1=Y_1=Y_2$是一个可能情况）。

这每个事件的概率的计算都不难，但是很繁．特别是总体分布有一个跃度为1/3的跳跃点$\frac{1}{2}$．例如，$P(A_1)$的计算要分解成以下5个部分：(a) Y_1或Y_2为$\frac{1}{2}$，其他全小于$\frac{1}{2}$；(b) 所有样本都小于$\frac{1}{2}$．(c) Y_1，Y_2中有一个大于$\frac{1}{2}$，一个为$\frac{1}{2}$；(d) Y_1,Y_2中有一个大于$\frac{1}{2}$，一个小于$\frac{1}{2}$；(e) Y_1，Y_2都大于$\frac{1}{2}$．分别计算这5种情况的概率再相加。

4-12 把平方展开逐项计算．结果为$m-1$，与χ^2_{m-1}的期望相同．这样做是希望统计量的确切分布与其极限分布能更接近一些。

4-14 注意到V_{jn}之值只涉及到以（X_{j1}，…，X_{jn_j}）为一方，以（X_{i1}，…，X_{in_i}，$i=1$，…，$j-1$）为另一方之间，每一对值（各方出一个）的大小比较，因此，（V_{2n}，…，$V_{r-1,n}$）只涉及（X_{i1}，…，X_{in_i}，$i=1$，…，$r-1$）的值内部大小比较问题，即它们的排列次序．由于全部样本为iid.，这个内部排列次序不影响以其整体为一方，以（X_{r1}，…，X_{rn_r}）为另一方的值的大小比较．由此就推出V_{rn}与（V_{2n}，…，$V_{r-1,n}$）独立（也不难写出形式的证明），由此推出V_{2n}，…，V_{mn}相互独立。

因为 $R(n)=V_{2n}+\cdots+V_{mn}$，右边各项独立，且当 $n\to\infty$ 时，每个 V_{in} 为渐近正态，故 $R(n)$ 也为渐近正态。

4-17 易见：n 个随机向量 $(X'_{1j}, \cdots, X'_{mj}): j=1, \cdots, n$ 为 iid．故若以 R_j 记集① $\{R_{1j}, \cdots, R_{mj}\}$，$j=1, \cdots, n$，则无论怎样把 1，2，…，$mn$ 分成 n 堆 $S_1, \cdots, S_n$，每堆 S_j 包含 m 个数，则 $P(R_j=S_j, j=1, \cdots, n)$ 与堆的分法无关（且就等于把 mn 个数分成 n 堆，每堆有 m 个数的不同分法的倒数，即 $(m!)^n/(mn)!$）、另一方面，对固定的 j，$(X'_{1j}, \cdots, X'_{mj})$ 为"可交换的"，即不论怎样作置换成 $(X'_{i_1j}, \cdots, X'_{i_mj})$，分布不变。由此知，在已知 $R_j=S_j$ 的条件下，$m!$ 种可能排列为等概率的这两个事实结合即证明了所要的结果。

4-21 以 $X_{(1)}\leqslant\cdots\leqslant X_{(n)}$ 记次序统计量，先利用分布函数与经验分布函数的非降与右连续性证明

$$\sup_{-\infty<x<\infty}|F(x)-F_n(x)|$$

$$=\max\left\{\left|F(X_{(i)})-\frac{i-1}{n}\right|, \left|F(X_{(i)})-\frac{i}{n}\right|: i=1, \cdots, n\right\}$$

再利用定理 2·1 即得．

第 五 章

5-1 必须的条件是：当原假设成立时，全部样本为独立同分布，或至少为"可交换的"（即各变量作置换不影响分布），而在对立假设下则没有这个性质。例 5·2 的模型 1 为独立同分布情况，而模型 2 为"可交换"情况，"对称中心为 0"的检验问题之所以不能用置换检验，是因为无论在原假设或对立假设之下，样本都是独立同分布。

---

① 注意这里讲的是集而非向量，即不计其中元素的次序，下面的"堆"$S_1, \cdots, S_n$ 也都是指集。$R_j=S_j$ 是在集合相等的意义下去理解。

5-2 下面是可考虑的检验之一：以 $\hat{m}$ 记合样本中位数，$X_i'=X_i-\hat{m}$，$Y_j'=Y_j-\hat{m}$．对（X_1'，…，X_{n_1}'，Y_1'，…，Y_{n_2}'）施行置换，并取统计量T为$\sum_{j=1}^{n_2}|Y_j'|$．请证明：在原假设成立之下，（X_1'，…，X_{n_1}'，Y_1'，…，Y_{n_2}'）的分布为可交换的，因而置换检验可用．又解释一下取统计量 T 的理由。

5-3 用归纳法证本题，首先易通过使用归纳假设，将问题转化为 $l=2$ 的情况．因为，若 $l=2$ 时已证，则把 $B_1,\cdots,B_{l-1}$ 结合为一新水平 B_1'（这时 B 有两个水平 B_1' 和 B_l），将得出：在给定行列和的条件下，（X_{1l}，…，X_{kl}）之条件分布与$\{p_i，q_j\}$无关．但是，在给定行列和的条件下再给定 X_{1l}，…，X_{kl}，等于在 $k\times$（$l-1$）列联表中给定行列和，按归纳假设，这时$\{X_{ij}:i=1，\cdots，k-1，j=1，\cdots，l-2\}$之条件分布与$\{p_i，q_j\}$无关。二者结合，即推出当（$k$，$l$-1）成立时对（$k,l$）也成立。同法由（$k-1$，$l$）推出（$k$，$l$）．

就 $l=2$的情况证本题，再一次使用归纳法，这次是对 k．首先注意：当 $l=2$ 时，若给定行列和，则 X_{11} 的条件分布与$\{p_i，q_j\}$无关。此因若把 A_2，…，A_k 结合为一新水平 A_1' 则回到例 5·1 已处理过的情况．这一点证明后即可在给定 X_{11} 后对 k 施行归纳法。

此法说理繁一些，但不需任何计算。

5-4 易见 $P(L_n=0)=\binom{n-[n^{1/5}]}{[n^{1/5}]}\Big/\binom{n}{[n^{1/5}]}\to 1$，当 $n\to\infty$．因 $E(L_n)=l_n=[n^{1/5}]^2/n$，

$$\sigma_n^2=\text{Var}(L_n)=\frac{1}{n-1}\left(\frac{[n^{1/5}](n-[n^{1/5}])}{n}\right)^2,$$

易见 $P((L_n-l_n)/\sigma_n=\sqrt{n-1}[n^{1/5}]/n)\to 1$，

而$\sqrt{n-1}[n^{1/5}]/n\to 0$，以此知（$L_n-l_n$）/$\sigma_n$不可能依分布收敛于 N（0，1）（事实上，它收敛到退化于一点 0 的分布）。

第 六 章

6-9 使用8题可知$\lim\limits_{n\to\infty}\sup\limits_{x}|Ef_n(x)-g(x)|=0$. 再用反证法，假设存在$x_0$使$F(x_0)-F(x_0-)>0$. 往证存在常数$c$及$x_n$使$Ef_n(x_n)\geqslant\frac{1}{h_n}K(c)[F(x_0)-F(x_0-)]$，导致矛盾。

6-17 利用以下事实：设Q为任一非空集合，f、g为定义在Q上的两个实函数，则有

$$|\sup_Q f(x)-\sup_Q g(x)|\leqslant\sup_Q|f(x)-g(x)|.$$

6-18 (1) $\delta^*(x)=E(Y|x)=\frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}x+\frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}\mu$,

$$R^*=\sigma_1^2\sigma_2^2/(\sigma_1^2+\sigma_2^2)$$

(2) $$r_n(x,x_1',\cdots x_k')=\left(\frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}\right)^2(x-x_n^*)^2+\left(1+\frac{1}{k}\right)\sigma_1^2\sigma_2^2/(\sigma_1^2+\sigma_2^2),$$

其中$X_n^*=\frac{1}{k}\sum_{i=1}^{k}x_i'$.

6-20 对充分大的$a>0$，存在与a无关的常数$c>0$使$P(|x-X_n'|\geqslant a|X=x)\leqslant c\left(\frac{1}{a-x}\right)^n$. 从而证明$\{x-X_n',\ n=1,2,\cdots\}$是一致可积。

符号与名词术语

一、与样本有关的

设$X_1,\cdots,X_n$为自一具分布F的总体中抽出的样本。若$X_1,\cdots,X_n$独立同分布（有时简记为iid.）且有公共分布F，则$X_1,\cdots,X_n$称为自该总体或分布F中抽出的简单样本。有时

把这记为 $X_1, \cdots, X_n \sim F$。

二、与分布有关的

分布函数总取为右连续的，即 X 的分布函数定义为 $P(X\leqslant x)$.一维与多维正态分布如常记为$N(\mu, \sigma^2)$与$N(a, A)$，维数略去不计。(a, b)区间上的均匀分布记为$R(a, b)$.自由度为 n 的 t 分布、χ^2 分布，以及自由度为(m, n)的 F 分布，如平常分别记为t_n, χ_n^2与F_{mn}. 这些分布的"上α分位点"($0<\alpha<1$)分别记为$t_n(\alpha)$, $\chi_n^2(\alpha)$与$F_{mn}(\alpha)$. 例如，$t_n(\alpha)$的意义是：$P(t_n>t_n(\alpha))=\alpha$.标准正态分布 $N(0, 1)$ 的上 α 分位点记为 μ_α. X, Y 同分布记为$X\overset{d}{=}Y$。

三、与数字特征有关的

随机变量 X 的(数学)期望及方差分别记为$E(X)$及$\mathrm{Var}(X)$(括号有时省略).两个随机变量 X,Y 的协方差记为$\mathrm{Cov}(X,Y)$。随机向量 X 的协方差阵记为$\mathrm{Cov}(X)$. 随机变量 X 在给定$Y=y$的条件下的条件期望记为$E(X|Y=y)$. 在不致引起误会时也记为$E(X|y)$，这样，$E(X|Y)$应理解为$E(X|y)|_{y=Y}$。

四、与随机变量(向量)及分布的收敛有关的

一串随机变量$\{X_n\}$依概率或以概率1收敛于一随机变量X，分别记为$X_n\xrightarrow{P}X$ 或 $X_n\longrightarrow X$, a.s.. 后者也记为$\lim\limits_{n\to\infty}X_n=X$, a.s.. 若$F_n$和$F$分别为$X_n$及$X$的分布，而对$F$的每个连续点$x$有$\lim\limits_{n\to\infty}F_n(x)=F(x)$，称$\{x_n\}$依分布收敛于变量$X$或分布$F$. 也称分布序列$\{F_n\}$依分布收敛于$X$或$F$，记为$F_n\xrightarrow{\mathscr{L}}F$等等。

五、与线代数或其他有关的

当提到一向量a时总是指列向量，a'为a的转置，故a'为行向量。矩阵用一个字母(例如A)记。m行n列的矩阵，其(i, j)元为a_{ij}者，有时记为$(a_{ij})_{i=1,\cdots,m,\ j=1,\cdots,n}$，向量

$a=(a_1, \cdots, a_n)'$ 的欧氏长度，即$\left(\sum_{i=1}^{n}a_i^2\right)^{1/2}$（取正号），记为 $\|a\|$. 单位方阵记为 I. 集（合）A 中所含元素个数记为 $\#(A)$。集 A 的指示函数，即在 A 上为 1 而在其外为 0 的函数，记为 $I(A)$.

$b_n=o(a_n)$ 和 $b_n=O(a_n)$ 分别表示 "$\lim_{n\to\infty}(b_n/a_n)=0$" 及 "$\{b_n/a_n: n=1, 2, \cdots\}$ 为有界序列"。类似记号也用于连续变量的情况。

sup 和 inf 分别表示上确界和下确界。

其他偶尔用到的符号与名词术语，将在用到的地方指明其意义。